Optimal clustering under uncertainty

Lori A. Dalton[1]*, **Marco E. Benalcázar**[2], **Edward R. Dougherty**[3]

1 Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, United States of America, **2** Departamento de Informática y Ciencias de la Computación, Facultad de Ingeniería de Sistemas, Escuela Politécnica Nacional, Quito, Ecuador, **3** Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, United States of America

* dalton.lori@outlook.com

Abstract

Classical clustering algorithms typically either lack an underlying probability framework to make them predictive or focus on parameter estimation rather than defining and minimizing a notion of error. Recent work addresses these issues by developing a probabilistic framework based on the theory of random labeled point processes and characterizing a *Bayes clusterer* that minimizes the number of misclustered points. The Bayes clusterer is analogous to the Bayes classifier. Whereas determining a Bayes classifier requires full knowledge of the feature-label distribution, deriving a Bayes clusterer requires full knowledge of the point process. When uncertain of the point process, one would like to find a robust clusterer that is optimal over the uncertainty, just as one may find optimal robust classifiers with uncertain feature-label distributions. Herein, we derive an optimal robust clusterer by first finding an *effective* random point process that incorporates all randomness within its own probabilistic structure and from which a Bayes clusterer can be derived that provides an optimal robust clusterer relative to the uncertainty. This is analogous to the use of effective class-conditional distributions in robust classification. After evaluating the performance of robust clusterers in synthetic mixtures of Gaussians models, we apply the framework to granular imaging, where we make use of the asymptotic granulometric moment theory for granular images to relate robust clustering theory to the application.

Introduction

The basic optimization paradigm for operator design consists of four parts: (1) define the underlying random process; (2) define the class of potential operators; (3) characterize operator performance via a cost function; and (4) find an operator to minimize the cost function. The classic example is the Wiener filter, where the four parts consist of wide-sense stationary true and observed signals, linear operators, minimization of the mean-square error, and optimization in terms of power spectra. In practice, we might be uncertain as to the distribution governing the random process so that we desire a *robust* operator, one whose performance is acceptable relative to the uncertainty. Robust design can be posed in the following way: Given a class of operators and given that the *state of nature* is uncertain but contained in some

uncertainty class, which operator should be selected to optimize performance across all possible states of nature?

Our interest here is in clustering, where the underlying process is a random point set and the aim is to partition the point set into clusters corresponding to the manner in which the points have been generated by the underlying process. Having developed the theory of optimal clustering in the context of random labeled point sets where optimality is with respect to mis-clustered points [1], we now consider optimal clustering when the underlying random labeled point process belongs to an uncertainty class of random labeled point processes, so that optimization is relative to both clustering error and model uncertainty. This is analogous to finding an optimal Wiener filter when the signal process is unknown, so that the power spectra belong to an uncertainty class [2]. We now briefly review classical robust operator theory, which will serve as the foundation for a new general theory of optimal robust clustering.

Optimal robust filtering first appeared in signal processing in the 1970s when the problem was addressed for signals with uncertain power spectra. Early work considered robust filter design from a minimax perspective: the filter is designed for the state having the best worst-case performance over all states [3–5]. Whereas the standard optimization problem given certainty with regard to the random process takes the form

$$\psi^* = \arg\min_{\psi \in \mathcal{C}} \gamma(\psi),$$ (1)

where $\mathcal{C}$ is the operator class and $\gamma(\psi)$ is the cost of applying operator ψ on the model, minimax optimization is defined by

$$\psi_{\mathrm{MM}} = \arg\min_{\psi \in \mathcal{C}_\Theta} \max_{\theta \in \Theta} \gamma_\theta(\psi),$$ (2)

where Θ is the uncertainty class of random processes, $\mathcal{C}_\Theta$ is the class of operators that are optimal for some state in the uncertainty class, and $\gamma_\theta(\psi)$ is the cost of applying operator ψ for state $\theta \in \Theta$.

Suppose one has prior knowledge with which to construct a prior distribution $\pi(\theta)$ on states (models) in the uncertainty class. Rather than apply a minimax robust operator, whose average performance can be poor, a Bayesian approach can be taken whereby optimization is relative to $\pi(\theta)$. A *model-constrained (state-constrained) Bayesian robust (MCBR) operator* minimizes the expected error over the uncertainty class among all operators in $\mathcal{C}_\Theta$:

$$\psi_{\mathrm{MCBR}} = \arg\min_{\psi \in \mathcal{C}_\Theta} E_\theta[\gamma_\theta(\psi)].$$ (3)

MCBR filtering has been considered for morphological [6], binary [7] and linear [8] filtering. MCBR design has also been applied in classification with uncertain feature-label distributions [9].

Rather than restrict optimization to operators that are optimal for some state in the uncertainty class, one can optimize over any class of operators, including unconstrained optimization over all possible measurable functions. In this case, the optimal operator is called an *intrinsically optimal Bayesian robust (IBR) operator (filter)* and Eq (3) becomes

$$\psi_{\mathrm{IBR}} = \arg\min_{\psi \in \mathcal{C}} E_\theta[\gamma_\theta(\psi)],$$ (4)

where $\mathcal{C}$ is a set of operators under consideration. IBR filtering has been considered for linear and morphological filtering [2]. The IBR approach was first used to design optimal classifiers when the unknown true feature-label distribution belongs to an uncertainty class [10, 11]. In

that setting, optimization is relative to a posterior distribution obtained from the prior utilizing sample data and an optimal classifier is called an *optimal Bayesian classifier (OBC)*.

Unlike the state of affairs in filtering and classification, classical clustering algorithms typically lack an underlying probability framework to make them predictive. The exceptions, for instance, expectation-maximization based on mixture models, typically focus on parameter estimation rather than defining and minimizing a notion of operator error. Work in [12] and [1] addresses the solution to Eq (1) in the context of clustering using a probabilistic theory of clustering for random labeled point sets and a definition of clustering error given by the expected number of "misclustered" points. This results in a *Bayes clusterer*, which minimizes error under the assumed probabilistic framework. An (optimal) Bayes clusterer is analogous to an (optimal) *Bayes* classifier, which minimizes classification error under the assumed feature-label distribution. Here, we characterize robust clustering using the framework and definitions of error in [12] and [1], and introduce definitions of robust clustering that parallel concepts from filtering. In particular, we present minimax, MCBR and IBR clusterers, and develop effective stochastic processes for robust clustering. We also evaluate performance under mixtures of Gaussians and demonstrate how the methodology can be implemented in practice with an example from granular imaging.

Bayes clustering theory

In this section, we review Bayes clustering theory from [12] and [1]. A *random labeled point process* (RLPP) is characterized by a pair, (Ξ, Λ), where Ξ is a point process generating a point set $S \subset \mathbb{R}^d$ and Λ generates random labels on the points in S. In particular, let $\eta(S)$ denote the number of points in S. The first component in this pair, Ξ, maps from a probability space to $(\mathbf{N}, \mathcal{N})$, where $\mathbf{N}$ is the family of finite sequences in $\mathbb{R}^d$ and $\mathcal{N}$ is the smallest σ-algebra on $\mathbf{N}$ such that for any Borel set B in $\mathbb{R}^d$ the mapping $S \mapsto \eta(S \cap B)$ is measurable. A probability measure, ν, of Ξ is determined by the probabilities $\nu(Y)$ for $Y \in \mathcal{N}$, or (via the Choquet-Matheron-Kendall theorem [13–16]), may be reduced to the system of probabilities $P(\Xi \cap K \neq \emptyset)$ over all compact sets $K \subseteq \mathbb{R}^d$. Given a point set $S \in \mathbf{N}$, a label function $\phi_S: S \to L = \{1, 2, \ldots, l\}$ is a deterministic mapping that assigns each point $\mathbf{x} \in S$ to label $\phi_S(\mathbf{x}) \in L$. The second component, Λ, is a random labeling, that is, $\Lambda = \{\Phi_S: S \in \mathbf{N}\}$, where Φ_S is a random label function with probability mass $P(\Phi_S = \phi_S | S)$ on L^S.

For any set S, and pair of label functions ϕ_S and φ_S, define the *label mismatch error* between ϕ_S and φ_S to be the proportion of points where the label functions differ:

$$\varepsilon(S, \phi_S, \varphi_S) = \frac{1}{\eta(S)} \sum_{\mathbf{x} \in S} I_{\phi_S(\mathbf{x}) \neq \varphi_S(\mathbf{x})}, \tag{5}$$

where I_A is an indicator function equal to 1 if A is true and 0 otherwise. Clustering involves identifying partitions of a point set rather than the actual labeling. A partition of S into l clusters has the form $\mathcal{P}_S = \{S_1, S_2, \ldots, S_l\}$ such that the S_y are disjoint and $S = \bigcup_{y=1}^{l} S_y$. Every partition $\mathcal{P}_S$ has associated with it a family, $G_{\mathcal{P}_S}$, of label functions that induce the partition $\mathcal{P}_S$. That is, $\varphi_S \in G_{\mathcal{P}_S}$ if and only if $\mathcal{P}_S = \{S_1, S_2, \ldots, S_l\}$ where $S_y = \{\mathbf{x} \in S: \varphi_S(\mathbf{x}) = \ell_y\}$ and $(\ell_1, \ldots, \ell_l)$ is a permutation of L. For any point set S, label function ϕ_S, and partition $\mathcal{P}_S$, define the *cluster mismatch error* to be the minimum label mismatch error between ϕ_S and all label functions that induce $\mathcal{P}_S$:

$$\varepsilon(S, \phi_S, \mathcal{P}_S) = \min_{\varphi_S \in G_{\mathcal{P}_S}} \varepsilon(S, \phi_S, \varphi_S). \tag{6}$$

This is a simplified version of the original definition in [12]. Define the *partition error* of $\mathcal{P}_S$ to be the mean cluster mismatch error over the distribution of label functions on S:

$$
\begin{aligned}
\varepsilon(S, \mathcal{P}_S) &= E_{\Phi_S}[\varepsilon(S, \Phi_S, \mathcal{P}_S)|S] \\
&= E_{\Phi_S}\left[\min_{\varphi_S \in G_{\mathcal{P}_S}} \varepsilon(S, \Phi_S, \varphi_S)|S\right].
\end{aligned}
\tag{7}
$$

In [1], it was shown that Eq (7) can be written in the form

$$
\varepsilon(S, \mathcal{P}_S) = \sum_{\mathcal{Q}_S \in \mathcal{K}_S} c_S(\mathcal{P}_S, \mathcal{Q}_S) P_S(\mathcal{Q}_S),
\tag{8}
$$

where $\mathcal{K}_S$ is the set of all partitions of S,

$$
P_S(\mathcal{Q}_S) = \sum_{\phi_S \in G_{\mathcal{Q}_S}} P(\Phi_S = \phi_S|S)
\tag{9}
$$

is the probability mass function on partitions $\mathcal{Q}_S \in \mathcal{K}_S$ of S, and we define the *natural partition cost function*,

$$
c_S(\mathcal{P}_S, \mathcal{Q}_S) = \frac{1}{\eta(S)} \min_{\varphi_S \in G_{\mathcal{P}_S}, \phi_S \in G_{\mathcal{Q}_S}} \sum_{\mathbf{x} \in S} I_{\phi_S(\mathbf{x}) \neq \varphi_S(\mathbf{x})}.
\tag{10}
$$

The partition error under the natural cost function is essentially the average number of misclustered points.

Taking Eq (8) as a generalized definition, other cost functions can be applied [17–20]. The natural cost function stands out in two respects. First, while these works define loss over label functions, we define cost directly over partitions, which is mathematically cleaner, and automatically treats the label switching problem in which multiple distinct label functions may produce the same partitions. Second, these works treat loss abstractly without connecting to a practical notion of clustering error, like the expected (minimum) number of mislabeled points. In contrast, we begin with a practical definition of clustering error, and prove that this error is equivalent to the average cost given in Eq (8) under the natural cost function.

A cluster operator ζ maps point sets to partitions. Define the *clustering error* of cluster operator ζ to be the mean partition error of $\zeta(\Xi)$ over the random point sets Ξ:

$$
\varepsilon(\zeta) = E_{\Xi}[\varepsilon(\Xi, \zeta(\Xi))].
$$

A *Bayes cluster operator* ζ^* is a clusterer having minimal clustering error $\varepsilon(\zeta^*)$, which is called the *Bayes clustering error*. Since $\varepsilon(S, \zeta(S))$ depends on the clusterer ζ only at point set S, $\varepsilon(\zeta)$ is minimized by setting $\zeta^*(S) = \mathcal{P}_S^*$ for all $S \in \mathbf{N}$, where $\mathcal{P}_S^*$ is a *Bayes partition* of S, defined to be a partition having minimal partition error, $\varepsilon(S, \mathcal{P}_S^*)$, called the *Bayes partition error*.

This formulation parallels classification theory, where an RLPP corresponds to a feature-label distribution, $\varepsilon(S, \mathcal{P}_S)$ corresponds to the probability that a given label is incorrect for a fixed point in the feature space, $\varepsilon(\zeta)$ corresponds to the overall classification error for an arbitrary classifier, ζ^* corresponds to a Bayes classifier, and $\varepsilon(\zeta^*)$ corresponds to the Bayes classification error.

To find the Bayes partition, we must evaluate the partition error for all partitions. We call partitions with non-zero probability *reference partitions*, and denote a set of r reference partitions by $\mathcal{R}_S = \{\mathcal{Q}_S^1, \ldots, \mathcal{Q}_S^r\} \subseteq \mathcal{K}_S$. We call partitions that comprise the search space *candidate partitions*, and denote a set of c candidate partitions by $\mathcal{C}_S = \{\mathcal{P}_S^1, \ldots, \mathcal{P}_S^c\} \subseteq \mathcal{K}_S$. The set of candidate partitions is not required to contain all reference partitions, and may even contain

non-reference partitions. The partition error of all candidate partitions is:

$$
\begin{bmatrix}
\varepsilon(S, \mathcal{P}_S^1) \\
\vdots \\
\varepsilon(S, \mathcal{P}_S^c)
\end{bmatrix}
=
\begin{bmatrix}
c_S(\mathcal{Q}_S^1, \mathcal{P}_S^1) & \cdots & c_S(\mathcal{Q}_S^r, \mathcal{P}_S^1) \\
\vdots & \ddots & \vdots \\
c_S(\mathcal{Q}_S^1, \mathcal{P}_S^c) & \cdots & c_S(\mathcal{Q}_S^r, \mathcal{P}_S^c)
\end{bmatrix}
\begin{bmatrix}
P_S(\mathcal{Q}_S^1) \\
\vdots \\
P_S(\mathcal{Q}_S^r)
\end{bmatrix}.
\tag{11}
$$

If $\mathcal{C}_S = \mathcal{R}_S = \mathcal{K}_S$, then we require a cost matrix of size $|\mathcal{K}_S| \times |\mathcal{K}_S|$, which can be prohibitively large for moderate $\eta(S)$. To alleviate this, [1] provides both exact and approximate techniques to evaluate Eq (11) under the natural cost function with reduced complexity.

Separable RLPPs

Up to this point, we have characterized RLPPs with a point process Ξ that generates point sets, S, followed by an S-conditioned labeling process Λ that generates label functions, ϕ_S. Alternatively, it is often easier to characterize an RLPP as a process that draws a sample size n, a set of labels for n points, and a set of n points with distributions corresponding to the labels. For instance, one might think of points being drawn from l Gaussian distributions possessing random parameters. We say that an RLPP is *separable* if a label function ϕ is generated from an independent label generating process Φ with probability mass function $P(\Phi = \phi)$ over the set of all label functions with domain $\{1, 2, \ldots, n\}$, a random parameter vector ρ is independently drawn from a distribution $f(\rho)$, and the ith point $\mathbf{x}_i$ in S, with corresponding label $y = \phi(i)$, is independently drawn from a conditional distribution $f(\mathbf{x}|y, \rho)$. From Bayes' rule, the probability of label function $\phi_S \in L^S$ given $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is

$$
P(\Phi_S = \phi_S | S) \propto f(S|\phi) P(\Phi = \phi),
\tag{12}
$$

where $\phi(i) = \phi_S(\mathbf{x}_i)$,

$$
f(S|\phi) = \int \left(\prod_{y=1}^{l} \prod_{\mathbf{x} \in S_y} f(\mathbf{x}|y, \rho) \right) f(\rho) d\rho,
\tag{13}
$$

and $S_y = \{\mathbf{x}_i : \phi(i) = y, i = 1, \ldots, n\}$ is the set of points in S assigned label y. A separable RLPP thus has three components: $P(\Phi = \phi)$, $f(\rho)$ and $f(\mathbf{x}|y, \rho)$, where $P(\Phi = \phi)$ is a prior on labels, which is not dependent on S, and $P(\Phi_S = \phi_S | S)$ is a posterior probability on labels given a specific point set S, which is found using Eqs (12) and (13).

If $\rho = [\rho_1, \ldots, \rho_l]$, where the ρ_y are mutually independent parameter vectors and the label-y-conditional distribution depends on only ρ_y, that is, if $f(\mathbf{x}|y, \rho) = f(\mathbf{x}|y, \rho_y)$ for $y = 1, \ldots, l$, then,

$$
f(S|\phi) = \prod_{y=1}^{l} \int \left(\prod_{\mathbf{x} \in S_y} f(\mathbf{x}|y, \rho_y) \right) f(\rho_y) d\rho_y.
\tag{14}
$$

Gaussian RLPPs

Expressions for label function probabilities have been solved under several models in [1]. Here, we review an important case in which clusters are Gaussian with random means and covariances. Specifically, consider a separable RLPP where, for each $y \in \{1, \ldots, l\}$, $\rho_y = [\mu_y, \Sigma_y]$ and $f(\mathbf{x}|y, \rho_y)$ is a Gaussian distribution with mean μ_y and covariance Σ_y. Given a label function ϕ_S, let $y \in \{1, \ldots, l\}$ be fixed, and let n_y be the number of points in S assigned label y. For $n_y \geq 2$

it was shown in [1] that

$$\prod_{\mathbf{x} \in S_y} f(\mathbf{x}|y, \rho_y) = (2\pi)^{-\frac{dn_y}{2}} |\Sigma_y|^{-\frac{n_y}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Phi_y^* \Sigma_y^{-1}\right)\right), \tag{15}$$

where

$$\Phi_y^* = (n_y - 1)\widehat{\Sigma}_y + n_y(\mu_y - \widehat{\mu}_y)(\mu_y - \widehat{\mu}_y)^T,$$

$\widehat{\mu}_y$ and $\widehat{\Sigma}_y$ are the sample mean and covariance of points in S_y, respectively, and where $|\cdot|$ denotes a determinant, $tr(\cdot)$ a trace and T a transpose. When $n_y = 1$, Eq (15) holds with $\Phi_y^* = (\mu_y - \widehat{\mu}_y)(\mu_y - \widehat{\mu}_y)^T$, and when $n_y = 0$ the product over an empty set is 1.

Assume $f(\rho_y) = f(\Sigma_y)f(\mu_y|\Sigma_y)$, where $f(\mu_y|\Sigma_y)$ is a Gaussian distribution with mean $\mathbf{m}_y$ and covariance $\frac{1}{v_y}\Sigma_y$ with $v_y > 0$, and $f(\Sigma_y)$ is an inverse-Wishart distribution with $\kappa_y > d - 1$ degrees of freedom and a positive-definite scale matrix Ψ_y, i.e.,

$$f(\Sigma_y) = \frac{|\Psi_y|^{\frac{\kappa_y}{2}}|\Sigma_y|^{-\frac{\kappa_y+d+1}{2}}}{2^{\frac{\kappa_y d}{2}}\Gamma_d\left(\frac{\kappa_y}{2}\right)} \exp\left(-\frac{1}{2}\mathrm{tr}(\Psi_y\Sigma_y^{-1})\right),$$

where Γ_d is the multivariate Gamma function. The expected mean is $\mathbf{m}_y$, the expected covariance matrix is $\frac{1}{\kappa_y - d - 1}\Psi_y$ if $\kappa_y > d + 1$, and as v_y and κ_y increase $f(\rho_y)$ becomes more "informative." The probability of label function ϕ_S under this RLPP is found from Eqs (12) and (14) as

$$P(\Phi_S = \phi_S|S) \propto P(\Phi = \phi) \prod_{y=1}^{l} \frac{\Gamma_d\left(\frac{\kappa_y+n_y}{2}\right)}{|n_y + v_y|^{\frac{d}{2}}|\Psi_y + \Psi_y^*|^{\frac{\kappa_y+n_y}{2}}}, \tag{16}$$

where

$$\Psi_y^* = (n_y - 1)\widehat{\Sigma}_y + \frac{v_y n_y}{v_y + n_y}(\widehat{\mu}_y - \mathbf{m}_y)(\widehat{\mu}_y - \mathbf{m}_y)^T \tag{17}$$

for $n_y = 2$, $\Psi_y^* = \frac{v_y}{v_y+1}(\widehat{\mu}_y - \mathbf{m}_y)(\widehat{\mu}_y - \mathbf{m}_y)^T$ for $n_y = 1$, and $\Psi_y^* = 0$ for $n_y = 0$. If $v_1 = \cdots = v_l$, $\kappa_1 = \cdots = \kappa_l$ and $P(\Phi = \phi)$ is such that the size of each cluster is fixed and partitions with clusters of the specified sizes are equally likely, then for any ϕ_S inducing clusters of the correct sizes,

$$P(\Phi_S = \phi_S|S) \propto \prod_{y=1}^{l} |\Psi_y + \Psi_y^*|^{-\frac{\kappa_y+n_y}{2}}. \tag{18}$$

Similar derivations for the posterior on parameters under Gaussian mixture models can be found in [21], and similar derivations for the posterior on label functions under Gaussian mixture models can be found in [17].

Robust clustering operators

Under a known RLPP (Ξ, Λ), optimization in the Bayes clusterer is over the set $\bar{\mathcal{C}}$ of all clustering algorithms with respect to the clustering error,

$$\zeta^* = \arg\min_{\zeta \in \bar{\mathcal{C}}} \varepsilon(\zeta); \tag{19}$$

however, in practice the RLPP is likely to be uncertain. In this section we present definitions for optimal Bayesian robust clustering and show that IBR clusterers solve an optimization problem of the same form as in Eq (19) under an effective process.

Definitions of robust clustering

We present three robust clustering operators: minimax robust clustering, model-constrained Bayesian robust (MCBR) clustering, and intrinsically optimal Bayesian robust (IBR) clustering. Our main interest is in IBR clustering. The first two methods are provided to emphasize parallels between the new theory and existing robust operator theory from filtering and classification.

Consider a parameterized uncertainty class of RLPPs $(\Xi_\theta, \Lambda_\theta)$, $\theta \in \Theta$, where Ξ_θ is a point process on $(\mathbf{N}, \mathcal{N})$, $\Lambda_\theta = \{\Phi_{\theta, S} : S \in \mathbf{N}\}$ is a random labeling on $\mathbf{N}$ consisting of a random label function $\Phi_{\theta, S}$ for each S, and $\varepsilon_\theta(\zeta)$ is the error of cluster operator ζ for state θ.

A *minimax robust clusterer* ζ_{MM} is defined by Eq (2) with $\mathcal{C}_\Theta$ being the set of state-specific Bayes clusterers and $\varepsilon_\theta(\zeta)$ in place of $\gamma_\theta(\psi)$. An *MCBR cluster operator* ζ_{MCBR} is defined by Eq (3) with $\varepsilon_\theta(\zeta)$ in place of $\gamma_\theta(\psi)$.

Our focus is on optimization over the full class $\bar{\mathcal{C}}$ of cluster operators. This yields an *IBR cluster operator*,

$$\zeta_{\mathrm{IBR}} = \arg\min_{\zeta \in \bar{\mathcal{C}}} E_\theta[\varepsilon_\theta(\zeta)]. \tag{20}$$

In analogy to [2], where effective characteristics for IBR linear filtering were derived from effective random signal processes, here we show how IBR cluster operators can be found via effective random labeled point processes.

Effective random labeled point processes

We begin with two definitions.

Definition 1. *An RLPP is solvable under clusterer class $\mathcal{C}$ if*

$$\zeta^* = \arg\min_{\zeta \in \mathcal{C}} \varepsilon(\zeta)$$

can be solved under this process.

Definition 2. *Let Θ be an uncertainty class of RLPPs having prior $\pi(\theta)$. An RLPP $(\Xi_{\mathrm{eff}}, \Lambda_{\mathrm{eff}})$ is an effective RLPP under clusterer class $\mathcal{C}$ if for all $\zeta \in \mathcal{C}$ both the expected clustering error $E_\theta[\varepsilon_\theta(\zeta)]$ under the uncertainty class of RLPPs and the clustering error $\varepsilon_{\mathrm{eff}}(\zeta)$ under $(\Xi_{\mathrm{eff}}, \Lambda_{\mathrm{eff}})$ exist and*

$$E_\theta[\varepsilon_\theta(\zeta)] = \varepsilon_{\mathrm{eff}}(\zeta). \tag{21}$$

Theorem 1. *Let Θ parameterize an uncertainty class of RLPPs with prior $\pi(\theta)$. If there exists a solvable effective RLPP $(\Xi_{\mathrm{eff}}, \Lambda_{\mathrm{eff}})$ under clusterer class $\mathcal{C}$ with optimal clusterer ζ^*_{eff}, then $\zeta^*_{\mathrm{eff}} = \arg\min_{\zeta \in \mathcal{C}} E_\theta[\varepsilon_\theta(\zeta)]$. If $\mathcal{C} = \mathcal{C}_\Theta$, then $\zeta^*_{\mathrm{MCBR}} = \zeta^*_{\mathrm{eff}}$, and if $\mathcal{C} = \bar{\mathcal{C}}$, then $\zeta^*_{\mathrm{IBR}} = \zeta^*_{\mathrm{eff}}$.*

Proof. The proof is immediate from the definition of an effective RLPP and Eq (19):

$$\arg\min_{\zeta \in \mathcal{C}} E_\theta[\varepsilon_\theta(\zeta)] = \arg\min_{\zeta \in \mathcal{C}} \varepsilon_{\mathrm{eff}}(\zeta) = \zeta^*_{\mathrm{eff}}.$$

The solutions for MCBR and IBR clustering follow from their definitions.

To find an MCBR or IBR clusterer, we first seek an effective RLPP. This effective RLPP is not required to be a member of the uncertainty class parameterized by θ, but must be solvable. If $(\Xi_{\mathrm{eff}}, \Lambda_{\mathrm{eff}})$ is an effective RLPP under clusterer class $\mathcal{C}$, then it is an effective RLPP under any

smaller clusterer class. Hence, an effective RLPP found for IBR clustering is also an effective RLPP for MCBR clustering. However, not only are IBR clusterers better performing than MCBR clusterers across the uncertainty class, they are typically much easier to find analytically. In particular, the IBR clusterer is directly solved by importing methods from Bayes clustering theory, i.e., one may solve Eq (19) by minimizing the partition error over all partitions of a point set S under the effective RLPP. The MCBR clusterer, on the other hand, is significantly hampered by computational overhead in finding $\mathcal{C}_\Theta$ and actually evaluating the clustering error for each $\zeta \in C_\Theta$. The next theorem addresses the existence of effective RLPPs.

Theorem 2. *Let Θ parameterize an uncertainty class $\{(\Xi_\theta, \Lambda_\theta)\}_{\theta \in \Theta}$ of RLPPs with prior $\pi(\theta)$. There exists an RLPP, $(\Xi_{\text{eff}}, \Lambda_{\text{eff}})$, such that*

$$E_\theta[E_{\Xi_\theta, \Lambda_\theta}[g(\Xi_\theta, \Phi_{\theta, \Xi_\theta})|\theta]] = E_{\Xi_{\text{eff}}, \Lambda_{\text{eff}}}[g(\Xi_{\text{eff}}, \Phi_{\text{eff}, \Xi_{\text{eff}}})] \tag{22}$$

for any real-valued measurable function, g.

Proof. Suppose that the parameter θ is a realization of a random vector, $\vartheta : (\Omega, \mathcal{A}, P) \to (\Theta, \mathcal{B})$. Then $\{\vartheta^{-1}(\theta): \theta \in \Theta\}$ partitions the sample space, Ω. The point process Ξ_θ is thus a mapping

$$\Xi_\theta : (\vartheta^{-1}(\theta), \mathcal{A} \cap \vartheta^{-1}(\theta), P_\theta) \to (\mathbf{N}, \mathcal{N}),$$

where P_θ is the conditional probability and we assume $v_\theta(Y) = P_\theta(\Xi_\theta^{-1}(Y))$ for all $Y \in \mathcal{N}$ is known. Write the random labeling as $\Lambda_\theta = \{\Phi_{\theta, S}: S \in \mathbf{N}\}$, where $\Phi_{\theta,S}$ has a probability mass function $P(\Phi_{\theta, S} = \phi_S|\theta, S)$ on L^S. Given any real-valued measurable function g mapping from point set and label function pairs, let $X = g(\Xi, \Phi_\Xi)$ be a random variable where (Ξ, Φ_Ξ) is drawn from $\{(\Xi_\theta, \Lambda_\theta)\}_{\theta \in \Theta}$ with prior $\pi(\theta)$, and note $E_\theta[E[X|\theta]] = E[X]$.

Let $\Xi_{\text{eff}} : (\Omega, \mathcal{A}, P) \to (\mathbf{N}, \mathcal{N})$ be a mapping, where given a fixed $\omega \in \Omega$ we have a corresponding fixed realization $\theta = \vartheta(\omega)$ and we define $\Xi_{\text{eff}}(\omega) = \Xi_\theta(\omega)$. Note that

$$v(Y) \equiv P(\Xi_{\text{eff}}^{-1}(Y)) = E_\theta[v_\theta(Y)]$$

and Ξ_{eff} is a random point process. Define $\Lambda_{\text{eff}} = \{\Phi_{\text{eff},S}: S \in \mathbf{N}\}$, where $\Phi_{\text{eff},S}$ has a probability mass function

$$P(\Phi_{\text{eff},S} = \phi_S|S) = E_\theta[P(\Phi_{\theta,S} = \phi_S|\theta, S)]$$

for all $\phi_S \in L^S$. Thus, Λ_{eff} is a random labeling. Let $Z = g(\Xi_{\text{eff}}, \Phi_{\text{eff}, \Xi_{\text{eff}}})$ be a random variable where $(\Xi_{\text{eff}}, \Phi_{\text{eff}, \Xi_{\text{eff}}})$ is drawn from the RLPP we have constructed, $(\Xi_{\text{eff}}, \Lambda_{\text{eff}})$, and note $E[X] = E[Z]$.

Theorem 2 applies for any function $g(S, \phi_S)$, including the cluster mismatch error $g(S, \phi_S) = \varepsilon(S, \phi_S, \zeta(S))$, for any clusterer $\zeta \in \bar{\mathcal{C}}$. Thus, Eq (22) implies

$$
\begin{aligned}
E_\theta[\varepsilon_\theta(\zeta)] &= E_\theta[E_{\Xi_\theta, \Lambda_\theta}[\varepsilon(\Xi_\theta, \Phi_{\theta, \Xi_\theta}, \zeta(\Xi_\theta))|\theta]] \\
&= E_{\Xi_{\text{eff}}, \Lambda_{\text{eff}}}[\varepsilon(\Xi_{\text{eff}}, \Phi_{\text{eff}, \Xi_{\text{eff}}}, \zeta(\Xi_{\text{eff}}))] = \varepsilon_{\text{eff}}(\zeta).
\end{aligned}
$$

Hence, $(\Xi_{\text{eff}}, \Lambda_{\text{eff}})$ is an effective RLPP on $\bar{\mathcal{C}}$, covering MCBR and IBR clusterers.

The following corollary shows that for separable RLPPs, the effective RLPP is also separable and aggregates uncertainty within and between models.

Corollary 1. *Let each RLPP in the uncertainty class be parameterized by ρ with prior density $f(\rho|\theta)$, let Φ be an independent labeling process with a probability mass $P(\Phi = \phi)$ that depends on neither θ nor ρ, and denote the conditional distribution of points by $f(\mathbf{x}|y, \rho, \theta)$. Then the effective RLPP is separable with parameter $[\theta, \rho]$, prior $f(\theta, \rho)$, an independent labeling process with probability mass $P(\Phi = \phi)$, and conditional distributions $f(\mathbf{x}|y, \rho, \theta)$.*

Proof. Let the number of points, *n*, and the label function $\phi: \{1, \ldots, n\} \to L^n$ be fixed. For a fixed *θ*, the effective random point process $\Xi_{\text{eff}}(\omega)$ is set equal to $\Xi_\theta(\omega)$. Equivalently, a realization of $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ under the effective RLPP is governed by the distribution

$$f(S|\phi) = \int \left(\int \left(\prod_{i=1}^{n} f(\mathbf{x}_i|\phi(i), \rho, \theta) \right) f(\rho|\theta) d\rho \right) \pi(\theta) d\theta.$$

This is equivalent to a separable random point process with parameter $[\theta, \rho]$, prior $f(\theta, \rho) = \pi(\theta)f(\rho|\theta)$ and conditional distributions $f(\mathbf{x}|y, \rho, \theta)$. Since the labeling process is independent, the full effective RLPP is the separable RLPP given in the statement of the corollary.

A graphical model of the uncertainty class of RLPPs assumed in Corollary 1 is provided in Fig 1. A general description of how the IBR cluster may be found follows.

1. We assume an uncertainty class of RLPPs of the form stated in Corollary 1 and illustrated in Fig 1. In particular, we require the sample size, *n*, prior $\pi(\theta)$, label process probability mass function $P(\Phi = \phi)$, parameter prior $f(\rho|\theta)$ and conditional density $f(\mathbf{x}|\rho, \theta)$. These components characterize our prior knowledge about the point set generating process, and ideally are constructed and validated using available scientific knowledge about the problem at hand.

2. By Corollary 1, the effective RLPP is found by merging uncertainty in the state (across RLPPs) and parameters (within RLPPs). In particular, the effective RLPP is characterized by the sample size *n*, label process probability mass function $P(\Phi = \phi)$, parameter prior $f(\theta, \rho)$ and density $f(\mathbf{x}|y, \rho, \theta)$.

3. By Theorem 1, the IBR cluster is the Bayes (optimal) cluster under the effective RLPP. Given point set *S*, the IBR cluster outputs the partition $\mathcal{P}_S$ corresponding to the minimal error $\varepsilon(S, \mathcal{P}_S)$ in Eq (11). The natural cost function c_S is a constant function given by Eq (10), the partition probabilities are given by Eq (9), and the label function probabilities $P(\Phi_S = \phi_S|S)$ under the effective (separable) RLPP are given by Eq (12) with the likelihood function in Eq (13) using $f(\theta, \rho)$ in place of $f(\rho)$ and $f(\mathbf{x}|y, \rho, \theta)$ in place of $f(\mathbf{x}|y, \rho)$. An algorithmic description of the IBR cluster under separable RLPPs is provided in Algorithm 1. Algorithm 1 is equivalent to the Bayes cluster under the effective RLPP, and several optimal and suboptimal methods to improve upon this Bayes clustering algorithm are provided in [1].
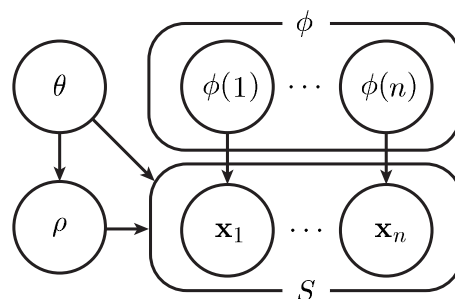

Fig 1. A graphical model of the uncertainty class of RLPPs assumed in Corollary 1. The parameter *θ* is governed by a prior distribution $\pi(\theta)$ and indexes each RLPP in the uncertainty class. The number of points, *n*, may be generated from an independent process, or considered fixed. For fixed *n*, the label function, *φ*, is generated according to the probability mass function $P(\Phi = \phi)$. Given *θ*, *ρ* is generated from the density $f(\rho|\theta)$, and each point in the point set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is drawn from the density $f(\mathbf{x}|y, \rho, \theta)$, where the corresponding label for point $\mathbf{x}_i$ is $y = \phi(i)$.

In practice, the primary issues are: (a) deriving an analytical form for the label function probability, $P(\Phi_S = \phi_S|S)$, (b) evaluating the natural cost, c_S, for all pairs of partitions, and (c) evaluating partition errors, $\varepsilon(S, \mathcal{P}_S)$, for all partitions. Note that $P(\Phi_S = \phi_S|S)$ is available for Gaussian separable RLPPs in Eq (16). Issues (b) and (c) may also be alleviated using optimal and suboptimal algorithms, as discussed in [1].

Algorithm 1 IBR Clustering for Separable RLPPs

```
Require: Data set, S
Require: Maximum number of clusters, l
Require: Label generating process, P(Φ = φ)
Require: Effective parameter generation process, f(θ, ρ)
Require: Effective data generation process, f(x|y, ρ, θ)
  n ← number of points in S
  K_S ← set of all possible partitions on n points with up to l clusters
  r ← number of partitions in K_S (number of reference partitions)
  c ← number of partitions in K_S (number of candidate partitions)
  normalize ← 0
  for i = 1 to r do
    Q_S ← K_S(i)
    for all label vectors φ that induce partition Q_S do
      f(S|φ) ← likelihood from Eq (13) using f(θ, ρ) and f(x|y, ρ, θ)
      a(φ) ← f(S|φ)P(Φ = φ)  (unnormalized label function prob. from Eq
      (12))
      normalize ← normalize + a(φ)
    end for
  end for
  for i = 1 to r do
    Q_S ← K_S(i)
    sum ← 0
    for all label vectors φ_S that induce partition Q_S do
      P(Φ_S = φ_S|S) ←a(φ)/normalize (normalized label function
      probability)
      sum ← sum + P(Φ_S = φ_S|S)
    end for
    P_S(i) ← sum (partition probability from Eq (9))
  end for
  for j = 1 to c do
    P_S ← K_S(j)
    for i = 1 to r do
      Q_S ← K_S(i)
      c_S(j, i) ← natural cost between P_S and Q_S from Eq (10)
    end for
    ε(S,j) ← Σ^r_{i=1} c_S(j,i)P_S(i) (clustering error from Eq (8))
  end for
  j* ← min_{j = 1, ..., c} ε(S, j)
  P*_S ← K_S(j*) (output the IBR partition)
```

Robust clustering under Gaussian RLPPs

Consider synthetic Gaussian data with $l = 2$ clusters in $d = 1, 2, 10, 100$ and $1,000$ dimensions. The state of nature is composed of the cluster covariances, and for a given state of nature the point process generates equal sized Gaussian clusters with random means and the corresponding covariances. Formally, we parameterize the uncertainty class of RLPPs with $\theta = [\theta_1, \theta_2]$, where $\theta_y = \Sigma_y$, and each Σ_y is drawn independently from an inverse-Wishart distribution with κ_y degrees of freedom and scale matrix Ψ_y. The RLPP in the uncertainty class corresponding to θ, $(\Xi_\theta, \Lambda_\theta)$, is a separable RLPP with parameter $\rho_y = \mu_y$, Gaussian prior $f(\rho_y|\theta_y)$ with mean

$\mathbf{m}_y$ and covariance Σ_y/v_y, and Gaussian conditional distributions $f(\mathbf{x}|y, \rho_y, \theta_y)$ with mean μ_y and covariance Σ_y. We set $\kappa_1 = \kappa_2 = d + 2$, $\Psi_1 = \Psi_2$ to be $d \times d$ identity matrices, $v_1 = v_2 = 1$, and $\mathbf{m}_1 = \mathbf{m}_2$ to be all-zero vectors. The number of points, $n = n_1 + n_2$, is set to 10 or 100 with $n_1 = n_2$, and the labels are permuted. Thus, the true distribution on label functions, $P(\Phi = \phi)$, has a support on the set of label functions that assign the correct number of points to each cluster, and is uniform on its support.

For each combination of d and n, we generate 1,000 states of nature, θ, and one point set per state of nature from the corresponding separable RLPP $(\Xi_\theta, \Lambda_\theta)$. For each point set, we run several classical clustering algorithms: fuzzy c-means (FCM), k-means (KM), hierarchical clustering with single linkage (H-S), hierarchical clustering with average linkage (H-A), hierarchical clustering with complete linkage (H-C), and a clusterer that produces a random partition with equal sized clusters for reference (Random). More details about these algorithms may be found in [22]. In addition, we cluster using expectation maximization for Gaussian mixture models (EM), and a method that minimizes a lower bound on the posterior expected variation of information under an estimated posterior similarity matrix generated from samples of a Gibbs sampler for Gaussian mixture models (MCMC) [23]. EM is run using the mclust package in R with default settings [24–26]. The Gibbs sampler is implemented using the bayesm package in R with 18,000 samples generated after a burn-in period of 2,000 samples, and otherwise default settings [27]. The posterior similarity matrix is estimated using the mcclust package in R [28], and minimization with respect to variation of information is implemented with the mcclust.ext package in R [29]. We also implement EM informed with the "correct" hyperparameters, κ_y, Ψ_y, v_y and $\mathbf{m}_y$ (EM-I) and MCMC informed with the "correct" hyperparameters (MCMC-I).

To find the IBR clusterer, the effective RLPP, (Ξ, Λ), is constructed using Corollary 1, which states that the effective RLPP merges uncertainty in the state θ with uncertainty in the parameter ρ. In this case, the effective RLPP is precisely the separable RLPP presented in the "Gaussian RLPPs" section, which accounts for both random means in ρ and random covariances in θ. The effective RLPP is solvable (at least for small point sets) using the Bayes clusterer presented in [1]. By Theorem 1, the IBR clusterer is equivalent to the Bayes clusterer under the effective RLPP. Thus, the IBR clusterer can be found when $n = 10$ by evaluating Eq (8) for all partitions using Eqs (9) and (18), and choosing the minimizing partition. When $n = 100$, we approximate the IBR clusterer (IBR-A) using a sub-optimal algorithm, Suboptimal Pseed Fast, presented in [1], which finds the maximum probability partition for a random subset of 10 points, generalizes these clusters to the full point set using a QDA classifier (in this case the threshold is selected such that $n_1 = n_2$), iteratively searches for the highest probability partition on the full point set by considering all partitions with at most two points clustered differently from the best partition found so far, and finally chooses the highest probability partition resulting from 10 repetitions with different initial subsets of points. MCBR and minimax robust clusterers are not found, since they are computationally infeasible. Furthermore, having found an IBR clusterer one would certainly not use an MCBR clusterer and very likely not use a minimax robust clusterer.

For each point set and each algorithm, we find the cluster mismatch error between the true partition and the algorithm output using Eq (6). For each combination of d and n and each algorithm, we approximate the average clustering error, $E_\theta[\varepsilon_\theta(\zeta)]$, under the natural cost function using the average cluster mismatch error across all 1,000 point sets. Fig 2A presents a graph of these errors with respect to d for $n = 10$, and similarly Fig 2B presents performance for $n = 100$. In part A the IBR clusterer is optimal. Although the IBR clusterer cannot be found in part B for $n = 100$, it is expected that its performance curve here should be slightly lower than its performance curve in part A for $n = 10$ [1]. Note the approximate IBR clusterer in part
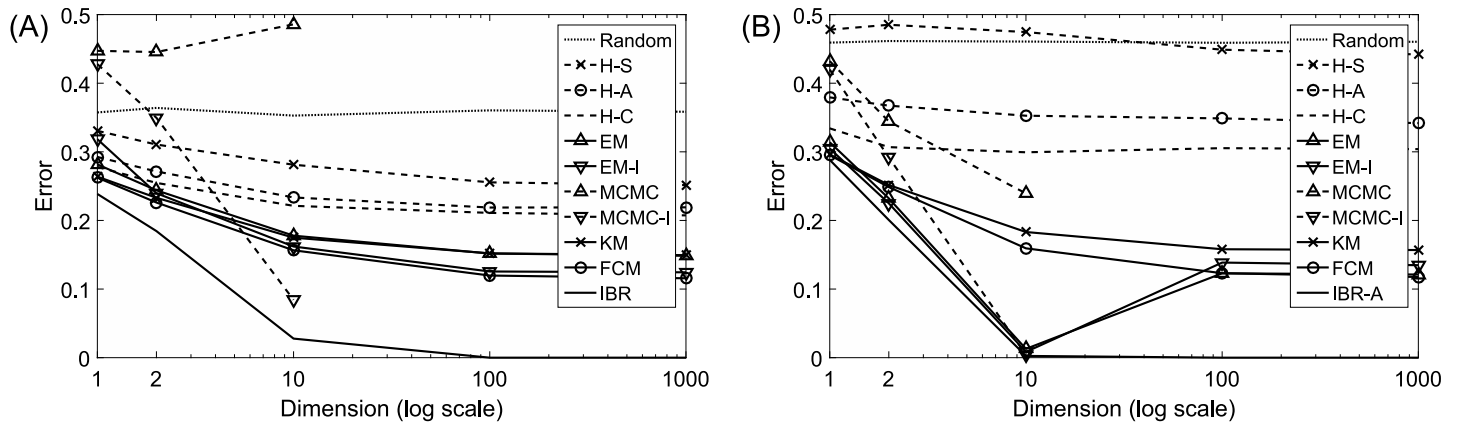
Fig 2. Average cluster mismatch error for Gaussian RLPPs. (A) $n = 10$. (B) $n = 100$.

B is slightly higher than the IBR clusterer in part A for 1 and 2 dimensions, slightly lower for 10 dimensions, and they both have nearly zero error for 100 and 1, 000 dimensions (in fact, they are the only algorithms with error rates below 10% in these high-dimension cases). This suggests that the IBR clusterer is very close to optimal, especially in easy problems where the error rate is very small.

When the number of points is large ($n = 100$) and the number of dimensions is smaller than the number of points, the performances of EM and EM-I are very close to the approximate IBR clusterer. However, when the number of points is small, or the number of dimensions is larger than the number of points, these algorithms tend to be similar to FCM and KM. This is most likely because mclust tests several different modeling assumptions regarding the covariances, and uses the Bayesian information criterion (BIC) to select a final output partition. When n is small relative to d, the full covariances of the Gaussian mixtures cannot be estimated well, so there is a tendency to select simpler models that assume covariances are equal and circular, which is essentially the same assumption made by FCM and KM. MCMC by default uses a particular normal-inverse-Wishart prior with hyperparameters that do not match the "correct" hyperparameters. The fact that MCMC-I performs much better than MCMC suggests that this method may be quite sensitive to the priors, especially when the sample size is small. Finally, note that MCMC and MCMC-I are not shown for $d = 100$ because they tend to be unstable in this case. For example, when $n = 100$ and $d = 100$ MCMC outputs all points in the same cluster for 913 out of 1, 000 iterations, MCMC-I outputs a label vector with the alternating pattern "1, 2, 1, 2, . . ." (i.e., it always assigns odd and even indexed points to separate clusters) for 849 out of 1, 000 iterations, and MCMC-I outputs this alternating pattern with at most three of the $n = 100$ points being exceptions for all of the remaining iterations. MCMC and MCMC-I are not shown for $d = 1, 000$ because the code crashes in this case.

Sample size does not influence the performance of IBR, IBR-A, KM or FCM very much, nor the performance of EM or EM-I under large dimensions. When dimensionality is small, EM and EM-I improve as we increase sample size, to the point where they approach the performance of the (nearly optimal) approximate IBR clusterer. In cases where MCMC and MCMC-I are working, they also improve as sample size increases, though to a lesser extent. One reason that EM and MCMC based methods may not perform particularly well under small samples is that they attempt to estimate the underlying Gaussian densities behind the clusters, and estimating an arbitrary covariance matrix becomes problematic as dimensionality increases. In contrast, the IBR clusterer does not attempt to estimate the densities, but rather

directly focuses on an easier problem: finding the best partition of the points. The performance of hierarchical methods all degrade as sample size increases, with H-S actually performing worse than random clustering in some cases. Hierarchical methods are sensitive to outliers and certain artifacts in the data that may be more likely to occur under large samples.

When the IBR cluster assumes $n_1 = n_2$, there are $r = C(n, n_1)/2$ reference partitions and $c = 2^{n-1}$ candidate partitions. For $n = 10$, $r = 126$ and $c = 512$ and the cost matrix in (11) contains $126 \times 512 = 64, 512$ elements. Although we use methods described in [1] to find the Bayes cluster efficiently without computing the whole cost matrix whenever possible (typically these tricks are most effective when the error rate is low), the size of the cost matrix increases very rapidly as n increases. The IBR cluster is currently infeasible to compute for more than about $n = 20$, which corresponds to $r = 92, 378$ reference and $c = 524, 288$ candidate partitions. Run times for algorithms run in the $n = 100$ case are shown in Table 1. Note we initialize IBR-A using the computationally intensive maximum probability partition for the subset of 10 points to improve error rates; when initializing with FCM run time improves at the expense of performance. An extensive analysis of the runtime and memory requirements for IBR and the FCM-based variant of IBR-A are provided in [1].

Robust clustering in granular imaging

While digital photography may now dominate over chemical photography, silver-based imaging remains important and is currently growing in use. Research remains active. Crystal shape is of particular importance. For many years granulometric analysis has been important in particle and texture analysis. In particular, morphological granulometries can generate image features relating to the size, shape, and concentration of particles. We present an application of robust clustering for images of silver-halide photographic T-grain crystals with respect to grain proportions using granulometric features.

Morphological granulometries

A basic model for silver-halide emulsions includes grains that are equilateral triangles, hexagons formed by removing triangle corners, rods (rectangles), and ill-formed blobs. To simplify calculations, we focus on a binary image model using only triangles and rods. In film grade emulsions grains overlap, but for laboratory analysis diluted emulsions with negligible overlapping can be produced, thus we also focus on images with non-overlapping grains.

Morphological granulometries are particularly well-suited for modeling and processing binary images consisting of grains of different sizes and shapes. The most commonly employed granulometry is a family of parameterized morphological openings: for a convex, compact *structuring element* (set) B, a *granulometry* $\{\Psi_t\}$ is defined by $\Psi_t(I) = I \circ tB$ for $t > 0$ and $\Psi_0(I) = I$, where $I \circ tB = \cup\{tB + x: tB + x \subset I\}$ is the opening of image (set) I by tB (more general granulometries exist [15]). If $\Omega_I(t)$ is the area of $\Psi_t(I)$, then $\Omega_I(t)$ is a decreasing function of t, known as a *size distribution*. A normalized size distribution is defined by $\Phi_I(t) = 1 - \Omega_I(t)/\Omega_I(0)$. If I is

Table 1. Average run time (in ms) over 1, 000 iterations in the Gaussian example with $n = 100$.

d	IBR-A	FCM	KM	MCMC	MCMC-I	EM	EM-I	H-C	H-A	H-S	Random
1	4, 266	1.2	3.4	12, 436	12, 047	5.3	4.6	1.1	1.1	1.3	0.5
2	17, 077	1.2	3.2	11, 461	11, 256	21.7	15.6	1.1	1.1	1.4	0.3
10	15, 992	1.5	3.3	12, 416	11, 965	405.3	18.0	1.1	1.1	1.4	0.3
100	172, 104	5.8	4.0	142, 888	147, 804	18.4	20.2	1.3	1.2	1.6	0.3
1,000	241, 375	26.9	10.3	–	–	171.2	191.2	2.9	2.8	3.2	0.4

compact and B consists of more than a single point, then $\Phi_I(t)$ increases from 0 to 1 and is continuous from the left. Thus, it defines a probability distribution function called the *pattern spectrum* of I (relative to B). Moments of $\Phi_I(t)$ are used for image classification and segmentation [30]. $\Phi_I(t)$ is a random function and its non-central moments (called *granulometric moments*) are random variables.

In this work, we use granulometric moments as features for clustering. Given a set I, we extract as features the first q granulometric moments of I generated by granulometries arising from p structuring elements $B_1, B_2, \ldots, B_p$, where we denote the kth granulometric moment corresponding to B_j by $\mu^{(k)}(I, B_j)$ for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, q$. Consider a random set I of the form

$$I = \bigcup_{i=1}^{m} \bigcup_{j=1}^{N_i} (r_{ij} A_i + x_{ij}),$$ (23)

where $A_1, A_2, \ldots, A_m$ are compact sets called *primitives*, r_{ij} and x_{ij} specify the radius (grain size) and center of the jth grain of primitive type i, respectively, and all $N = N_1 + \ldots + N_m$ grains are mutually disjoint.

In the silver halide application, we assume preprocessed images are well modeled by Eq (23), where $m = 2$, A_1 is an equilateral triangle with horizontal base, and A_2 is a vertical rod with height 5 times its base. Without loss of generality, we assume both primitives have unit area, i.e., $v[A_1] = v[A_2] = 1$, and we denote the grain proportions by b_1 and b_2. We further assume the r_{ij} are independent with the $r_{i1}, \ldots, r_{iN_i}$ identically distributed, where the *grain sizing distribution* for primitive i has the property $E[r_{ij}^k] = \gamma_{ik}\beta^k$ for all $k > 0$ and γ_{ik} and β are positive constants. If $r_{ij} \sim \text{gamma}(\alpha_i, \beta)$, β being the scale parameter for both primitives, then this property holds with $\gamma_{ik} = \Gamma(\alpha_i + k)/\Gamma(\alpha_i)$.

For the morphological opening, we use $p = 2$ structuring elements, where B_1 and B_2 are, respectively, vertical and horizontal linear structuring elements. The first $q = 2$ granulometric moments for B_1 and B_2 are

$$\mathbf{z} = \left[\mu^{(1)}(I, B_1) \; \mu^{(1)}(I, B_2) \; \mu^{(2)}(I, B_1) \; \mu^{(2)}(I, B_2) \right]^T.$$

Given the constants $\mu^{(k)}(A_i, B_j)$ and the radii r_{ij} of all grains, the exact moments in $\mathbf{z}$ under the granulometric model may be found analytically (see Theorem 1 in S1 Appendix). In particular, $\mathbf{z} = M\mathbf{x}$, where

$$M = \begin{bmatrix} \mu^{(1)}(A_1, B_1) & \mu^{(1)}(A_2, B_1) & 0 & 0 \\ \mu^{(1)}(A_1, B_2) & \mu^{(1)}(A_2, B_2) & 0 & 0 \\ 0 & 0 & \mu^{(2)}(A_1, B_1) & \mu^{(2)}(A_2, B_1) \\ 0 & 0 & \mu^{(2)}(A_1, B_2) & \mu^{(2)}(A_2, B_2) \end{bmatrix},$$

and $\mathbf{x} = [x_{11}, x_{21}, x_{12}, x_{22}]^T$, where

$$x_{ik} = \frac{\sum_{j=1}^{N_i} r_{ij}^{k+2}}{\sum_{j=1}^{N_1} r_{1j}^2 + \sum_{j=1}^{N_2} r_{2j}^2}.$$ (24)

In general, the constants $\mu^{(k)}(A_i, B_j)$ under convex grains can be found using theory from [31]. It can be shown that for triangle A_1 and vertical structuring element B_1 that $\mu^{(1)}(A_1, B_1) = 2 \cdot 3^{-3/4}$ and $\mu^{(2)}(A_1, B_1) = 2^{-1}3^{1/2}$. Similarly, for other combinations of primitives and structuring elements, $\mu^{(1)}(A_1, B_2) = 4 \cdot 3^{-5/4}$, $\mu^{(2)}(A_1, B_2) = 2 \cdot 3^{-1/2}$, $\mu^{(1)}(A_2, B_1) = 5^{1/2}$, $\mu^{(2)}(A_2, B_1) = 5$, $\mu^{(1)}(A_2, B_2) = 5^{-1/2}$ and $\mu^{(2)}(A_2, B_2) = 5^{-1}$.

In the current application, we cluster on the features $\mathbf{x} = M^{-1}\mathbf{z}$. Theorem 3 in S1 Appendix. guarantees asymptotic joint normality and provides analytic expressions for the asymptotic mean and covariance of granulometric moments under multiple primitives and multiple structuring elements. In particular, given the grain proportions b_1 and b_2, and the grain sizing parameters β and γ_{ik} for $i = 1, 2$ and $k = 2, 3, 4$, $\mathbf{x}$ has asymptotic mean

$$\frac{1}{b_1\gamma_{12} + b_2\gamma_{22}}\left[\begin{array}{cccc} b_1\gamma_{13}\beta & b_2\gamma_{23}\beta & b_1\gamma_{14}\beta^2 & b_2\gamma_{24}\beta^2 \end{array}\right]^T \tag{25}$$

and covariance matrix

$$\frac{1}{N(b_1\gamma_{12} + b_2\gamma_{22})^4}\begin{bmatrix} A_{11}\beta^2 & A_{12}\beta^3 \\ A_{21}\beta^3 & A_{22}\beta^4 \end{bmatrix}, \tag{26}$$

where the A_{ij} are 2×2 matrices that depend on only the b_i and γ_{ik}, and are provided in S1 Appendix.

Robust clustering

Suppose we are given a collection of n binary images of mixtures of silver-halide photographic T-grain crystals, where each image belongs to one of two groups, indexed by $y = 1, 2$. Images in class 1 and 2 have different proportions of triangles, b_1, and different sizing parameters, thereby providing different photographic properties. Our objective is to cluster the images into the two groups (our concern is partitioning, not labeling) based on feature vectors $\mathbf{x} = M^{-1}\mathbf{z}$ obtained from moments of morphological openings $\mathbf{z}$.

Given the grain sizing distributions and a prior $f(\rho)$, the asymptotic joint normality of $\mathbf{x}$ motivates a separable RLPP model where, given y and ρ, $f(\mathbf{x}|y, \rho)$ is a Gaussian distribution with mean and covariance given by Eqs (25) and (26), respectively. We substitute ρ and $1-\rho$ in place of b_1 and b_2 under class 1, and vice-versa under class 2. For simplicity, we assume $P(\Phi = \phi)$ is uniform with support such that the number of images in each class is known.

The grain sizing distribution in a binarized image typically depends on the image thresholding method and other factors, and thus is unknown. To account for this, we model an uncertainty class of RLPPs parameterized by θ, where the grain sizes are assumed to be gamma (α_{iy}, β_y) distributed, the α_{iy} parameters are fixed and known, the β_y depend deterministically on θ, and we assume θ and ρ are mutually independent with known prior $\pi(\theta)$. From Eqs (12) and (13), the IBR clusterer reduces to finding the following label function probabilities under the effective RLPP:

$$P(\Phi_S = \phi_s|S) \propto P(\Phi = \phi) \times$$
$$\int_0^\infty \int_0^1 f(S_1|1, \rho, \theta)f(S_2|2, \rho, \theta)f(\rho)\pi(\theta)d\rho d\theta, \tag{27}$$

where

$$f(S_y|y, \rho, \theta) = \prod_{\mathbf{x} \in S_y} f(\mathbf{x}|y, \rho, \theta).$$

Since we assume a Gaussian model, $f(S_y|y, \rho, \theta)$ is precisely the likelihood function in Eq (15). To make Eq (27) tractable, we assume discrete priors on ρ and θ so that the integrals can be written as sums.

Experimental results

The image generation model is based on the parameterized RLPP model described above. For a given set of images under a given RLPP with parameter θ, which determines the sizing distribution, we generate $n = n_1 + n_2$ binary images, where n_1 and n_2 denote the fixed number of images from class 1 and class 2, respectively. Each image contains 1,000 non-overlapping and vertically aligned grains (triangles and rods), and is 550×550 pixels. The prior $f(\rho)$ on the proportion of triangles for class 1 is uniform over 500 values from 0.45 to 0.55, and we assume the proportion of triangles for class 2 is $1 - \rho$. Fig 3 shows three example realizations of images with gamma($\alpha = 1.95, \beta = 2$) sizing distributions for the triangles and gamma($\alpha = 1.97, \beta = 2$) for the rods. Parts A, B, and C contain triangle proportions 0.45, 0.5, and 0.55, respectively.

The prior $\pi(\theta)$ is uniform over 10 values from 1.75 to 2. We assume gamma(α_{iy}, β_y) sizing distributions for primitive i under class y, where $\beta_1 = \theta$, $\beta_2 = 3.75 - \theta$. For triangles, $\alpha_{1y} = 1.95$ and 1.97 for class 1 and class 2, respectively, and for rods, $\alpha_{2y} = 1.97$ and 1.95 for class 1 and class 2, respectively. We generate 500 sets of images for each state, for a total of 5,000 sets of images. For each image, openings are found, followed by granulometric moments $\mathbf{z}$ from the openings, and finally a feature vector $\mathbf{x} = M^{-1}\mathbf{z}$. Fig 4 provides example scatter plots of all pairs of features extracted from 100 images. These images correspond to $\theta = 1.75$ and the 10 smallest values of ρ (between 0.45 and 0.452), with 5 images selected from each value of ρ and each group.

For each set of images, we run FCM, KM, H-S, H-A, H-C, EM, MCMC and Random. Note EM-I and MCMC-I, which use normal-inverse-Wishart priors on the mean and covariance pairs, are not sensible to run here since the model uncertainty on b_1, b_2 and β is not very compatible with this prior form. We also find the IBR partition using the Bayes partition for the effective RLPP, which merges uncertainty in θ and ρ. In particular, we compute the partition error for all partitions of the images from Eq (11), and choose the partition with minimal partition error. Note that Eq (11) is found using the natural cost function in Eq (10), and the posterior partition probabilities in Eq (9), which is based on posterior label function probabilities that may be computed exactly using a discretized version of Eq (27). Recall $f(\mathbf{x}|y, \rho, \theta)$ is assumed Gaussian with means given by Eq (25), covariances given by Eq (26), and appropriate values for b_1, b_2 and β depending on y, ρ and θ. It is possible to list all partitions and compute the partition errors exactly when $n = 10$ and $l = 2$. Again, we did not test MCBR and minimax robust clusterers owing to their high computational cost.

Fig 5 shows the approximate clustering error for all algorithms with respect to θ, computed using the average cluster mismatch error over 500 sets of images for each θ. Part A shows results when $n_1 = n_2 = 5$, and part B shows results when $n_1 = 6$ and $n_2 = 4$. In both parts A and
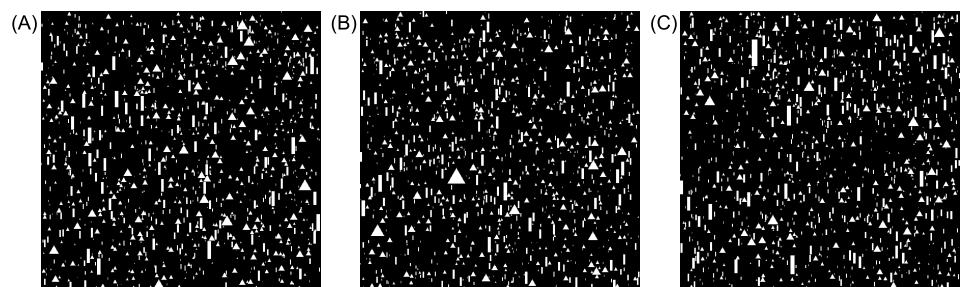

Fig 3. Examples of image realizations generated by the T-grain crystal model. Each image contains 1,000 grains. The sizing distribution of the grains are gamma($\alpha = 1.95, \beta = 2$) for the triangles and gamma($\alpha = 1.97, \beta = 2$) for the rods. The size of each image is 550×550 pixels. (A) Proportions of 0.45 triangles and 0.55 rods. (B) 0.5 triangles and 0.5 rods. (C) 0.55 triangles and 0.45 rods.
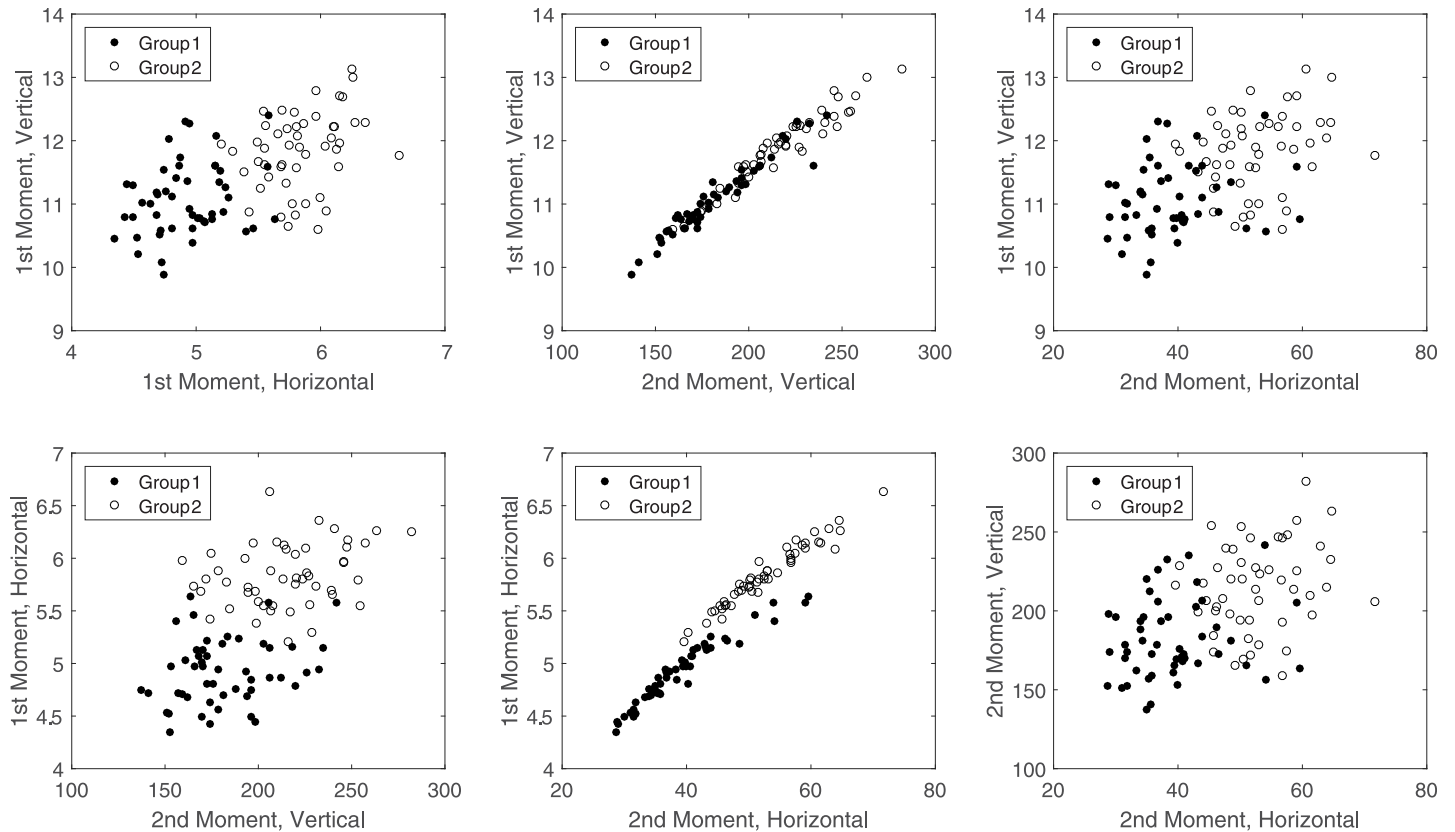
Fig 4. Scatter plots of all pairs of features extracted from 100 images.

B, the IBR clusterer performs much better than all classical algorithms across all states. Note that the IBR clusterer makes "incorrect" Gaussian modeling assumptions, but that the Gaussianity assumption and the analytically computed mean and covariance for each cluster become more accurate as the number of grains increases. Under all algorithms there is a significant variation in performance, which deteriorates when $\theta \approx 1.8750$. This corresponds to the
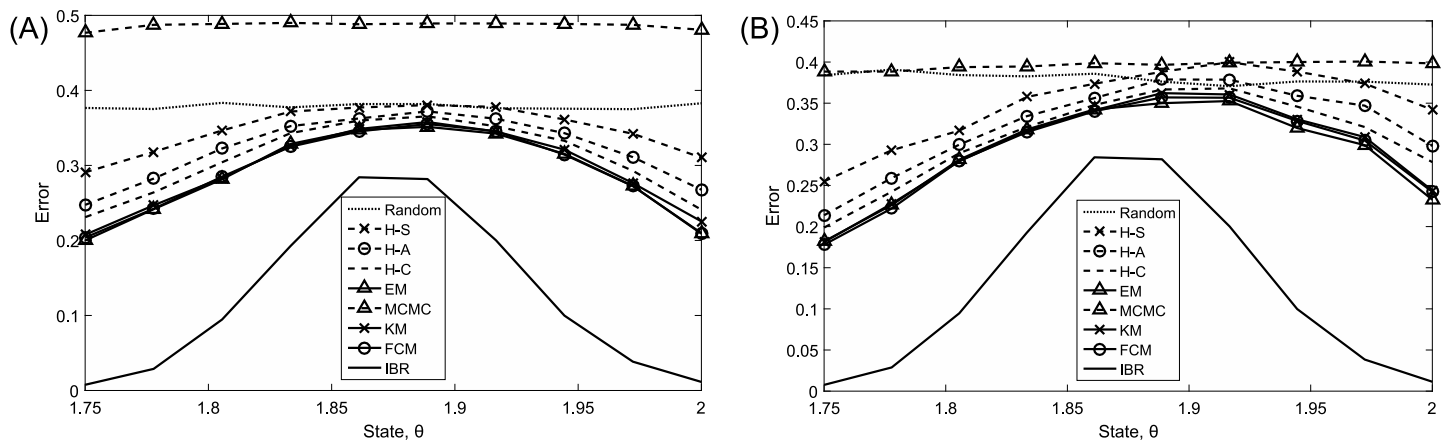

Fig 5. Average cluster mismatch error as a function of the state, θ, in the granular imaging example. (A) $n_1 = n_2 = 5$. (B) $n_1 = 6$ and $n_2 = 4$.

Table 2. Average cluster mismatch error over 5, 000 iterations in the granular imaging example.

n_1, n_2	IBR	FCM	KM	MCMC	EM	H-C	H-A	H-S	Random
5, 5	0.1239	0.2899	0.2938	0.4858	0.2890	0.3086	0.3223	0.3477	0.3786
6, 4	0.1351	0.2924	0.2952	0.3956	0.2904	0.3079	0.3224	0.3488	0.3799

case where $\beta_1 = \beta_2$, i.e., the case where the classes are most similar. Among all classical algorithms, the EM algorithm is usually the best, followed by FCM and KM, which have very similar performance. In some cases in Fig 5, the performance of hierarchical clustering with single linkage is worse than Random. As seen in the "Robust Clustering Under Gaussian RLPPs" section, MCMC with incorrect priors and small samples again has very poor performance. These graphs are summarized in Table 2, which shows the approximate average clustering error for each algorithm over all states and iterations. Finally, note that performance is similar between equal and unequal cluster size for all algorithms.

Since our focus is on robust clustering theory rather than image processing, in particular, the interplay between clustering optimization and the structure of prior knowledge, we have implemented a model setting based on Eq (23); nevertheless, before concluding this section, we believe a few comments concerning the effect of deviations from the model assumptions on the asymptotic granulometric moments are warranted.

The grain model of Eq (23) has been used in numerous studies of granulometric filtering and asymptotic moment analysis. Three issues regarding robustness of the theory to deviations from model assumptions have been addressed in [32]: (1) assuming a certain sizing distribution when in fact the random set satisfies a different sizing distribution, (2) using erroneous parameters for the sizing distribution, and (3) prior segmentation when there is modest overlapping.

For instance, the effect of erroneous gamma(α, β) sizing was analytically quantified with respect to misclassification error. Perhaps more importantly, the effect of watershed segmentation to separate overlapping grains prior to moment analysis was quantified by establishing lower and upper bounds on the actual kth granulometric moments when there are multiple grain primitives. One could reconsider the entire clustering analysis relative to these bounds; however, given the complexity of the bounds, this would involve a complicated mathematical study that would lead us far afield. The bounds are quite tight when grain overlapping is minor, as it is with a properly prepared emulsion.

Finally, as in all asymptotic granulometric theory, grain orientation is assumed fixed and not subject to rotation. The assumption is that each grain can be canonically rotated so that triangles have a horizontal base and for rods the shorter side forms the base, as assumed in the model. Robustness relative to imperfect rotation normalization has not been studied analytically. In fact, in digital image processing, rotation can cause problems for triangles and rectangles when edge detection is inaccurate, which is troublesome when there is low pixel resolution, a situation that is much less problematic today than when the basic granulometric theory was developed twenty years ago.

Conclusion

We have extended the theories of robust filtering and classification to clustering and developed new theory showing that optimal Bayesian robust clustering can be viewed as two equivalent optimization problems, one based on a parameterized uncertainty class of RLPPs and the other on a single effective RLPP that absorbs all parameters in the model. Thus, one can first

focus on modeling the uncertainties and then focus on finding the Bayes clusterer (or a good approximation) for the effective model.

The proposed paradigm for robust clustering is distinct from all other clustering methods in that it is fully model-based, can account for all prior knowledge and sources of uncertainty, and is optimal relative to clustering error. A key part of the paradigm involves justifying the modeling assumptions. In cases where the modeling assumptions can be justified, like in our granular imaging example where we developed new theory on the asymptotic joint normality and moments of our extracted features, we now have a very powerful theory for optimal robust clustering. Furthermore, since the Bayes and IBR clusterers employ powerful optimization directly with respect to clustering error (or clustering risk if used with specialized cost functions), under small to moderate imperfections of the assumed model they often continue to outperform many principled optimization-based methods. For instance, although our implementations of the EM, MCMC and IBR algorithms all assume Gaussian mixture models, EM and MCMC do not always perform as well as IBR because: (1) they focus on estimating the means and covariances instead of minimizing error, (2) they are often implemented without available prior knowledge.

IBR clustering is useful in a wide range of applications, particularly in clustering problems where the underlying data generation process is unknown, but can be theoretically constrained or partially described using scientific knowledge. Our granular imaging application is an excellent example, where we use a theorem that justifies Gaussianity and constrain parameters of the sizing distributions to train an IBR clusterer that far outperforms other data-driven methods. IBR clustering makes distributional assumptions that are expected to improve performance when correct, but may also degrade performance when grossly incorrect. Although we have defined and characterized IBR clustering for general applications here, in specific applications prior construction and prior validation are critically important steps.

The IBR clusterer (and Bayes clusterer) can only be implemented under small samples with up to 20 or so points; when n is large approximations of the IBR clusterer are available. Near optimal performance is possible with reasonable run time, and by tweaking the approximation algorithm one can make a trade-off between performance and run time. Suboptimal methods inspired by the optimal equations for the Bayes clusterer under Gaussian models (e.g., Suboptimal Pseed Fast) presented in [1] have good performance and competitive computation time with point sets of size up to 10, 000. Whether the exact or approximate IBR clusterer is used, increasing dimensionality also increases computation time, but to a much smaller extent; here we have implemented IBR clustering on datasets with up to 1, 000 features. Nevertheless, new suboptimal algorithms and methods to address computation remain important topics for further research. Since our objective here has been to develop a basic framework for robust clustering, we have focused on examples with relatively small point sets, implemented optimal algorithms whenever possible, and strongly favored better performing suboptimal algorithms at the expense of run time. We aim to continue developing improved algorithms for Bayes and IBR clustering in future work.

Supporting information

S1 File. Minimal data set for the Gaussian example. This file contains the number of mismatches (integer count of errors) between the correct label and label output by each algorithm (mismatch error is minimized over all sets of labels that induce the same partitions) in each iteration, and the run time for each algorithm in each iteration, which are used to generate Fig 2 and Table 1.
(ZIP)

S2 File. Minimal data set for the granular imaging example. This file contains the number of mismatches between the correct label and label output by each algorithm in each iteration, which are used to generate Fig 5 and Table 2. This file also contains raw data and MATLAB code used to generate Fig 4.
(ZIP)

S1 Appendix. Granulometry theorems. This file contains three granular imaging theorems that justify modeling assumptions like normality used by the IBR clusterer in The granular imaging example.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Lori A. Dalton, Edward R. Dougherty.

Formal analysis: Lori A. Dalton, Edward R. Dougherty.

Funding acquisition: Lori A. Dalton.

Investigation: Lori A. Dalton, Marco E. Benalcázar.

Methodology: Lori A. Dalton, Marco E. Benalcázar, Edward R. Dougherty.

Project administration: Lori A. Dalton, Edward R. Dougherty.

Software: Marco E. Benalcázar.

Supervision: Edward R. Dougherty.

Visualization: Marco E. Benalcázar.

Writing – original draft: Lori A. Dalton, Marco E. Benalcázar, Edward R. Dougherty.

Writing – review & editing: Lori A. Dalton.

References

1. Dalton LA, Benalcázar ME, Brun M, Dougherty ER. Analytic Representation of Bayes Labeling and Bayes Clustering Operators for Random Labeled Point Processes. IEEE Transactions on Signal Processing. 2015; 63(6):1605–1620. https://doi.org/10.1109/TSP.2015.2399870

2. Dalton L, Dougherty E. Intrinsically optimal Bayesian robust filtering. IEEE Transactions on Signal Processing. 2014; 62(3):657–670. https://doi.org/10.1109/TSP.2013.2291213

3. Kuznetsov VP. Stable detection when the signal and spectrum of normal noise are inaccurately known. Telecommunications and Radio Engineering. 1976; 30-31:58–64.

4. Kassam SA, Lim TI. Robust Wiener Filters. Journal of the Franklin Institute. 1977; 304(415):171–185. https://doi.org/10.1016/0016-0032(77)90011-4

5. Poor HV. On robust Wiener filtering. IEEE Trans Automatic Control. 1980; 25(4):531–536. https://doi.org/10.1109/TAC.1980.1102349

6. Dougherty ER, Chen Y. Robust Optimal Granulometric Bandpass Filters. Signal Processing. 2001; 81:1357–1372. https://doi.org/10.1016/S0165-1684(01)00017-2

7. Grigoryan AM, Dougherty ER. Design and analysis of robust binary filters in the context of a prior distribution for the states of nature. Mathematical Imaging and Vision. 1999; 11(3):239–254. https://doi.org/10.1023/A:1008356620614

8. Grigoryan AM, Dougherty ER. Bayesian Robust Optimal Linear Filters. Signal Processing. 2001; 81 (12):2503–2521. https://doi.org/10.1016/S0165-1684(01)00144-X

9. Dougherty ER, Hua J, Xiong Z, Chen Y. Optimal Robust Classifiers. Pattern Recognition. 2005; 38 (10):1520–1532. https://doi.org/10.1016/j.patcog.2005.01.019

10. Dalton LA, Dougherty ER. Optimal classifiers with minimum expected error within a Bayesian framework–Part I: Discrete and Gaussian models. Pattern Recognition. 2013; 46(5):1301–1314. https://doi.org/10.1016/j.patcog.2012.10.019

11. Dalton LA, Dougherty ER. Optimal classifiers with minimum expected error within a Bayesian framework–Part II: Properties and performance analysis. Pattern Recognition. 2013; 46(5):1288–1300. https://doi.org/10.1016/j.patcog.2012.10.019

12. Dougherty ER, Brun M. A Probabilistic Theory of Clustering. Pattern Recognition. 2004; 37(5):917–925. https://doi.org/10.1016/j.patcog.2003.10.003

13. Choquet G. Theory of capacities. Annales de l'institut Fourier. 1954; 5:131–295. https://doi.org/10.5802/aif.53

14. Kendall DG. Foundations of a theory of random sets. Stochastic Geometry. 1974; 3(9):322–376.

15. Matheron G. Random sets and integral geometry. New York: John Wiley & Sons; 1975.

16. Chiu SN, Stoyan D, Kendall WS, Mecke J. Stochastic geometry and its applications. 3rd ed. Wiley Series in Probability and Statistics. Chichester, West Sussex, United Kingdom: John Wiley & Sons; 2013.

17. Binder DA. Bayesian cluster analysis. Biometrika. 1978; 65(1):31–38. https://doi.org/10.1093/biomet/65.1.31

18. Quintana FA, Iglesias PL. Bayesian clustering and product partition models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2003; 65(2):557–574. https://doi.org/10.1111/1467-9868.00402

19. Fritsch A, Ickstadt K, et al. Improved criteria for clustering based on the posterior similarity matrix. Bayesian analysis. 2009; 4(2):367–391. https://doi.org/10.1214/09-BA414

20. Meilă M. Comparing clusterings—an information based distance. Journal of multivariate analysis. 2007; 98(5):873–895. https://doi.org/10.1016/j.jmva.2006.11.013

21. DeGroot MH. Optimal statistical decisions. Hoboken, NJ: John Wiley & Sons; 2005.

22. Dalton L, Ballarin V, Brun M. Clustering algorithms: On learning, validation, performance, and applications to genomics. Current genomics. 2009; 10(6):430–445. https://doi.org/10.2174/138920209789177601 PMID: 20190957

23. Wade S, Ghahramani Z. Bayesian cluster analysis: Point estimation and credible balls. arXiv preprint arXiv:150503339. 2015;.

24. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American statistical Association. 2002; 97(458):611–631. https://doi.org/10.1198/016214502760047131

25. Chris Fraley TBM Adrian E Raftery, Scrucca L. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Department of Statistics, University of Washington; 2012. Technical Report No. 597.

26. R Core Team. R: A Language and Environment for Statistical Computing; 2017. Available from: https://www.R-project.org/.

27. Rossi P. bayesm: Bayesian Inference for Marketing/Micro-Econometrics; 2017. Available from: https://CRAN.R-project.org/package=bayesm.

28. Fritsch A. mcclust: Process an MCMC Sample of Clusterings; 2012. Available from: https://CRAN.R-project.org/package=mcclust.

29. Wade S. mcclust.ext: Point estimation and credible balls for Bayesian cluster analysis; 2015.

30. Dougherty ER, Pelz J. Morphological Granulometric Analysis of Electrophotographic Images—Size Distribution Statistics For Process Control. Optical Engineering. 1991; 30:438–445. https://doi.org/10.1117/12.55823

31. Dougherty ER, Sand F. Representation of Linear Granulometric Moments for Deterministic and Random Binary Euclidean Images. Visual Communication and Image Representation. 1995; 6:69–79. https://doi.org/10.1006/jvci.1995.1006

32. Sand F, Dougherty ER. Robustness of Granulometric Moments. Pattern Recognition. 1999; 32 (9):1657–1665. https://doi.org/10.1016/S0031-3203(99)00028-X