TRACTABLE MEASURE OF COMPONENT OVERLAP FOR GAUSSIAN MIXTURE MODELS

EWA NOWAKOWSKA, JACEK KORONACKI, AND STAN LIPOVETSKY

Abstract. The ability to quantify distinctness of a cluster structure is fundamental for certain simulation studies, in particular for those comparing performance of different classification algorithms. The intrinsic integral measure based on the overlap of corresponding mixture components is often analytically intractable. This is also the case for Gaussian mixture models with unequal covariance matrices when space dimension $d > 1$. In this work we focus on Gaussian mixture models and at the sample level we assume the class assignments to be known. We derive a measure of component overlap based on eigenvalues of a generalized eigenproblem that represents Fisher's discriminant task. We explain rationale behind it and present simulation results that show how well it can reflect the behavior of the integral measure in its linear approximation. The analyzed coefficient possesses the advantage of being analytically tractable and numerically computable even in complex setups.

1. Introduction

1.1. **Overview.** There are numerous measures designed to capture distance between distributions or – more specifically – overlap between components of a Gaussian mixture model. One of the oldest is the Bhattacharyya coefficient (see for instance [1] or [2]), which reflects the amount of overlap between two statistical samples or distributions, a generalization of Mahalanobis distance described in [3] or [4]. In the context of information theory the most generic is the Kullback-Leibler divergence (see [5]) – a non-symmetric measure of difference between two distributions, also interpreted as expected discrimination information, which sets the link with possible classification performance. In [6] an overlap coefficient is proposed that measures agreement between two distributions, it is applied to samples of data coming from normal distributions. Among more recent works, in [7] a c-separation measure between multidimensional Gaussian distributions is defined, later developed in [8] as exact-c-separation. In [9], in the setup simplified to two clusters $k = 2$ and two dimensions $d = 2$, overlap rate is defined as a ratio of the joint density in its saddle point to its lower peak. The concept of ridge curve is introduced and further developed in [10] and [11], generalized to arbitrary number of dimensions and clusters, turning the ridge curve into a ridgeline manifold of the dimension $k - 1$.

All the measures use the parameters of the distributions to assess the overlap between the components and are typically formulated in terms of the underlying model. However, they can also be applied at the data level, as long as the class

2000 *Mathematics Subject Classification.* 62H30, 62E99.

Key words and phrases. mixture model, cluster structure, overlap measure.

Supported by National Science Center of Poland, DEC-2011/01/N/ST6/04174.

(or cluster) assignment is known. Then the model parameter estimates are used instead instead.

1.2. Content.

In Section 2 we recall the generic concept of component overlap and its best linear approximation, we also show an example of an overlap assessment approach and point to common difficulties. Then, in Section 3 we introduce what we refer to as Fisher's distinctness measure and we explain rationale behind it. Finally, in Section 4 we show results of a simulation study that illustrates how well the Fisher's coefficient can reflect the linear approximation of the original intractable overlap coefficient.

2. Overlap of Distributions

2.1. Integral measure.

The most generic and natural coefficient of overlap between components is what follows directly from the mixture definition:

$$\mathrm{MLE}_{\mathrm{err}} = 1 - \int_{\mathbb{R}^d} \max\big(\pi_1 f_1(\mu_1, \boldsymbol{\Sigma}_1), \ldots, \pi_k f_k(\mu_k, \boldsymbol{\Sigma}_k)\big)(x)\mathrm{d}x,$$

which for $k = 2$ classes simplifies to

$$\mathrm{MLE}_{\mathrm{err}} = \int_{\mathbb{R}^d} \min\big(\pi_1 f_1(\mu_1, \boldsymbol{\Sigma}_1), \pi_2 f_2(\mu_2, \boldsymbol{\Sigma}_2)\big)(x)\mathrm{d}x, \tag{1}$$

where for $d \geq 1$ by f_i, $i = 1, \ldots, k$ we denote component densities and by π_i, $i = 1, \ldots, k$ their corresponding mixing factors. Throughout this work we will assume though that equal mixing factors are assigned to all the components, which corresponds to balanced cluster sizes at the sample level. Coefficient (1) measures the actual overlap between two probability distribution and for $d = 1$ is illustrated in Figure 1. It coincides with intuitive understanding of components' overlap and with its expected behavior — grows with increasing within cluster dispersion (or variance, for $d = 1$) and decreasing distance between cluster centers. Also, it exhibits a strong link with classification performance, setting the upper limit for possible predictive accuracy in terms of maximum likelihood estimation (MLE) (see for instance [12]). Namely, best classification procedures based on likelihood ratio (MLE) or — equivalently — on its logarithm are given by

$$\mathrm{loglik}(f_1, f_2)(x) = \log\left(\frac{f_2(\mu_2, \boldsymbol{\Sigma}_2)(x)}{f_1(\mu_1, \boldsymbol{\Sigma}_1)(x)}\right). \tag{2}$$

For the value of (2) less than a constant observation x is classified to the first cluster, to the second otherwise. Hence the area of overlap between the components, as given by 1, corresponds to the expected proportion of observations that are incorrectly classified by MLE-classification rule, based on the (estimated) parameters of the mixture. Therefore (1) is denoted by $\mathrm{MLE}_{\mathrm{err}}$ and alternatively referred to as MLE-misclassification or error rate.

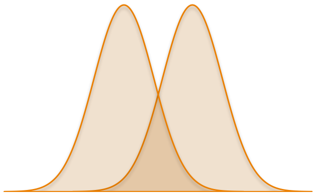

Figure 1. Overlap (dark shadow) between $k = 2$ Gaussian components in $d = 1$ dimension, concept illustration.

The fundamental problem with formula (1), and also one of the reasons for numerous alternative approaches to overlap assessment, is that (1) is hardly tractable for mixtures with different covariance matrices in higher dimensions. Handling it analytically would require integrating functions of Gaussian density over regions whose description often does not possess a tractable formulaic description either. Therefore, it can only be treated as a theoretical overlap coefficient for Gaussian mixture models and for practical applications replaced with other approaches.

2.2. **Best linear approximation.** The authors of [13] propose an approximation of (1) — best linear separator for $k = 2$ Gaussian components in $d \geq 1$ dimensions and an algorithm to determine it for a given data set X. They derive a linear function of $x \in \mathbb{R}^d$ given by a vector $b \in \mathbb{R}^d$ such that for a given constant $c \in \mathbb{R}$ inequality $b^T x \leq c$ classifies observation x to the first cluster, while $b^T x > c$ to the second. Vector b and constant c are obtained iteratively in order to minimize maximal probability of misclassification. As this approach will be used in our simulations, it is described below in more details following [13].

For x coming from component $l = 1, 2$, $b^T x$ has a univariate normal distribution with mean $b^T \mu_l$ and variance $b^T \mathbf{\Sigma}_l b$. As such, the probability of misclassifying observation x when it comes from the first population $l = 1$ equals

(3)
$$\mathbb{P}_1\left(b^T x > c\right) = \mathbb{P}_1\left(\frac{b^T x - b^T \mu_1}{b^T \mathbf{\Sigma}_1 b} > \frac{c - b^T \mu_1}{b^T \mathbf{\Sigma}_1 b}\right) = 1 - \Phi\left(\frac{c - b^T \mu_1}{b^T \mathbf{\Sigma}_1 b}\right) = 1 - \Phi(u_1),$$

where Φ denotes cumulative distribution function for a univariate standardized normal distribution (centered at zero, with variance equal to one) and $u_1 = \frac{c - b^T \mu_1}{b^T \mathbf{\Sigma}_1 b}$. Similarly, probability of misclassifying observation x when it comes from the second population $l = 2$ equals

$$\begin{aligned} \mathbb{P}_2\left(b^T x \leq c\right) = \mathbb{P}_1\left(\frac{b^T x - b^T \mu_2}{b^T \mathbf{\Sigma}_2 b} \leq \frac{c - b^T \mu_2}{b^T \mathbf{\Sigma}_2 b}\right) = \\ = \Phi\left(\frac{c - b^T \mu_2}{b^T \mathbf{\Sigma}_2 b}\right) = 1 - \Phi\left(\frac{b^T \mu_2 - c}{b^T \mathbf{\Sigma}_2 b}\right) = 1 - \Phi\left(u_2\right), \end{aligned} \tag{4}$$

for $u_2 = \frac{b^T \mu_2 - c}{b^T \mathbf{\Sigma}_2 b}$. As Φ is monotonic, the task

$$\max\left(\mathbb{P}_1(u_1), \mathbb{P}_2(u_2)\right) \longrightarrow \min_{\substack{b \in \mathbb{R}^d \\ c \in \mathbb{R}}}$$

is equivalent to

$$\min(u_1, u_2) \longrightarrow \max_{\substack{b \in \mathbb{R}^d \\ c \in \mathbb{R}}}, \tag{5}$$

which is more convenient to work with. As the objective is to find $b \in \mathbb{R}^d$ and $c \in \mathbb{R}$ that minimize maximal probability of misclassification, we will refer to the resulting procedure as a minimax procedure. Analytical formulation of admissible procedures for b and c leads to the following characterization

$$b = \left(t_1 \mathbf{\Sigma}_1 + t_2 \mathbf{\Sigma}_2\right)^{-1} (\mu_2 - \mu_1) \tag{6}$$

and

$$c = b^T \mu_1 + t_1 b^T \mathbf{\Sigma}_1 b = b^T \mu_2 - t_2 b^T \mathbf{\Sigma}_2 b, \tag{7}$$

where $t_1 \in \mathbb{R}$ and $t_2 \in \mathbb{R}$ are scalars. Minimax procedure is an admissible procedure with $u_1 = u_2$. As such, for $t = t_1 = (1 - t_2)$ the following equality must hold

$$0 = u_1^2 - u_2^2 = t^2 b^T \mathbf{\Sigma}_1 b - (1-t)^2 b^T \mathbf{\Sigma}_2 b = b^T \left[t^2 \mathbf{\Sigma}_1 - (1-t)^2 \mathbf{\Sigma}_2 \right] b. \tag{8}$$

Equation (8) for t can be solved numerically by means of iterative procedure.

With the above derivations, for a mixture of $k = 2$ components in $d \geq 1$ dimensions with parameters $\mu_1, \mathbf{\Sigma}_1$ and $\mu_2, \mathbf{\Sigma}_2$ respectively, the following algorithm provides best linear separator in terms of minimizing the maximal probability of misclassification.

Algorithm 2.1: BestLinearSeparator($\mu_1, \mathbf{\Sigma}_1, \mu_2, \mathbf{\Sigma}_2, prec$)

```
initialize $incr, crit, t$
repeat
  calculate $b$ with (6)
  calculate $crit$ with (8)
  if $crit > prec$
    then $t \leftarrow t - incr$
  if $crit < -prec$
    then $t \leftarrow t + incr$
  $incr \leftarrow incr \cdot \frac{1}{2}$
until criterion $crit$ given by (8) met with expected precision $prec$
calculate $c$ with (7)
calculate $u_1$ and $u_2$ and the probabilities of misclassification with (3) and (4)
calculate overall probability of misclassification $\mathbb{P}_{\text{minmax}} = \max(\mathbb{P}_1(u_1), \mathbb{P}_2(u_2))$
return $(\mathbb{P}_{\text{minmax}}, b, c, t)$
```

Note that the value of assumed precision *prec* must be given, while the values of scalar t, criterion *crit* and increment *incr* must be initialized. What is more, $\mathbb{P}_{\text{minmax}} = \mathbb{P}_1(u_1) = \mathbb{P}_2(u_2))$ as for the minimax procedure $u_1 = u_2$ must hold. Note, that $\mathbb{P}_{\text{minmax}}$ may be considered a measure of overlap as a result of linear approximation of criterion (2). If $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, formula (2) and its linear approximation given by b and c coincide, which is sure not the case for $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$.

2.3. **The challenge of replacement.** Degree of overlap between mixture components is critical for classification performance and must be assessed for simulation purposes and comparison of classification methods, hence the interest in the topic. There are many measures proposed in the literature that possess the property of being tractable even in a complex setup, however it is highly required that their behavior reflects the behavior of MLE_{err} based either on (1) or on its linear approximation of the previous subsection. However, this is not always the case, as shown in the below example.

E-distance. The method for overlap assessment proposed in [14] does not assume underling normal mixture model, however it can be very well applied in such setup. It is considered an extension to Ward's minimum variance method (see [15]) that formally takes both into account — heterogeneity between groups and homogeneity within groups in data. For this purpose it uses joint between-within e-distance between clusters that constitutes the basis for agglomerative hierarchical clustering procedure the authors propose. They define e-*distance* between two sets of observations $X_1 = \{x_{i_1} \colon c(i_1) = 1\}$, $n_1 = |X_1|$ and $X_2 = \{x_{i_2} \colon c(i_2) = 2\}$, $n_2 = |X_2|$ as

$$\text{(9)} \quad \mathrm{e}(X_1, X_2) = \frac{n_1 n_2}{n_1 + n_2} \left(\frac{2}{n_1 n_2} \sum_{i_1 \colon c(i_1)=1} \sum_{i_2 \colon c(i_2)=2} \|x_{i_1} - x_{i_2}\| + \right.$$
$$\left. - \frac{1}{n_1^2} \sum_{i_1 \colon c(i_1)=1} \sum_{j_1 \colon c(j_1)=1} \|x_{i_1} - x_{j_1}\| - \frac{1}{n_2^2} \sum_{i_2 \colon c(i_2)=2} \sum_{j_2 \colon c(j_2)=2} \|x_{i_2} - x_{j_2}\| \right).$$

The value of e-distance between two resulting clusters may be considered a cluster structure distinctness measure. It is expected to reflect changes in within-cluster dispersion and between-cluster separation. It should also remain in tune with the theoretical structure distinctness measure given by likelihood ratio (2) or its linear approximation from [13].

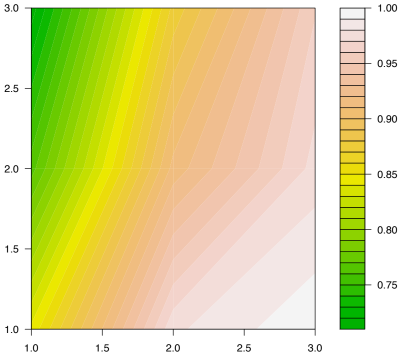

Figure 2. Heatmap — Anderson-Bahadur misclassification error w/r to growing between (x-axis) and within (y-axis) cluster dispersion.

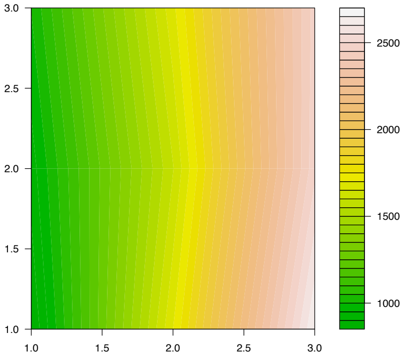

Figure 3. Heatmap — Székely-Rizzo e-distance (9) w/r to growing between (x-axis) and within (y-axis) cluster dispersion.

Figures 2 and 3 compare variability of structure distinctness measures based on Anderson-Bahadur ([13]) and Székely-Rizzo ([14]) proposals respectively. The former, similarly to the likelihood ratio theoretical distinctness measure, does depend on both — between-cluster distance and within-cluster dispersion, while the latter essentially remains insensitive to within cluster dispersion, depending entirely on the between class separation. This is an empirical insight which shows substantial discrepancy between behavior of theoretical and intuitive structure distinctness measure and e-distance given by (9), and hence points to another potential difficulty when trying to replace the integral coefficient.

3. Fisher's distinctness measure

3.1. **Model and notation.**

We consider a data set $X = (x_1, \ldots, x_n)^T,\ X \in \mathbb{R}^{n\times d}$ of n observations coming from a mixture of k d-dimensional normal distributions

$$f(x) = \pi_1 f_1(\mu_1, \mathbf{\Sigma}_1)(x) + \ldots + \pi_k f_k(\mu_k, \mathbf{\Sigma}_k)(x),$$

where

$$f_l(\mu_l, \mathbf{\Sigma}_l)(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det \mathbf{\Sigma}_l}} e^{-\frac{1}{2}(x-\mu_l)^T \mathbf{\Sigma}_l^{-1}(x-\mu_l)}.$$

We call each $f_l(\mu_l, \mathbf{\Sigma}_l)$, $l = 1, \ldots, k$ a component of the mixture and each π_l, $l = 1, \ldots, k$ a mixing factor of the corresponding component (see [12] or [16] and [17] or [18] for comparison with alternative approaches). We assume that for all the components equal mixing factors are assigned $\pi_1 = \cdots = \pi_k = \frac{1}{k}$. However, we allow different covariance matrices $\mathbf{\Sigma}_l$. Additionally, we assume large space dimension with respect to the number of components $d > k - 1$ and take the number of components k and class assignments as known.

We use lower index to indicate data set when sample estimates of parameters are used. In particular, by $\mu_X \in \mathbb{R}^d$ we denote sample mean and by $\mathbf{\Sigma}_X \in \mathbb{R}^{d\times d}$ covariance matrix for a data set X. For notation ease we center the data at the origin $\mu_X = 0$. We assume the covariance matrix to be of full rank, $\text{rank}(\mathbf{\Sigma}_X) = d$. Let $T_X = n\mathbf{\Sigma}_X$ be the total scatter matrix for X. We recall that a simple calculation (see for instance [12] or [19]) splits total scatter into its between and within cluster components $T_X = B_X + W_X$. By $\mu_{X,l}$ and $\mathbf{\Sigma}_{X,l}$ we denote empirical mean and covariance matrix for class l, where $l = 1, \ldots, k$. By $M_X = (\mu_{X,1}, \ldots, \mu_{X,k})$, $M_X \in \mathbb{R}^{d\times k}$ we understand a matrix of column vectors of cluster means. We assume the cluster means — as a set of points — to be linearly independent, so $\text{rank}(M_X) = \min(d, k-1) = k - 1$.

3.2. **Fisher's task as an eigenproblem.**

Originally (see [20]), the separation was defined for 2 classes in single dimension $v \in \mathbb{R}^d$ as the ratio of the variance between the classes to the variance within the classes

$$F_o(v) = \frac{v^T B_X v}{v^T W_X v}. \tag{10}$$

and then minimized over possible directions to find the linear subspace (Fisher's discriminant) that separates the classes best

$$v^* = \text{argmin}(F_0(v)).$$

For our purposes we will use the formulation

$$F(v) = \frac{1}{1 + \frac{1}{F_o(v)}} = \frac{v^T B_X v}{v^T T_X v}, \tag{11}$$

which is equivalent to (10) due to $T_X = B_X + W_X$ and yields the Fisher's subspace by maximizing over possible dimensions

$$v^* = \text{argmax}(F(v)). \tag{12}$$

As multiplying v by a constant does not change the result of (12), it can alternatively be expressed as a constrained optimization problem, namely

$$\begin{aligned} \max_{v\in\mathbb{R}^d} \quad & v^T B_X v \\ \text{subject to} \quad & v^T T_X v = 1. \end{aligned} \tag{13}$$

The corresponding Lagrange function defined as

$$L(v;\lambda) = v^T B_X v + \lambda\big(v^T T_X v - 1\big)$$

yields

$$\frac{\partial L(v;\lambda)}{\partial v} = 2B_X v - 2\lambda T_X v = 0,$$

so

$$B_X v = \lambda T_X v \tag{14}$$

must hold at the solution. Problem (14) is a generalized eigenproblem for two matrices B_X and T_X. As we assume covariance matrix to be well-defined, total scatter matrix T_X is invertible, however $T_X^{-1} B_X$ is not necessarily symmetric so it is a priori not obvious if the eigenvalues are real. Hence, a decomposition of the matrix T_X is required to reduce the generalized eigenproblem to a standard eigenproblem for a transformed matrix.

Solving a standard eigenproblem for T_X we obtain

$$T_X = A_{T_X} L_{T_X} A_{T_X}^T. \tag{15}$$

Note, that A_{T_X} is orthonormal (i.e. $A_{T_X} A_{T_X}^T = \mathbf{I}$ so $A_{T_X}^{-1} = A_{T_X}^T$). Replacing in (14) matrix T_X with its spectral decomposition (15) we get

$$B_X v = \lambda A_{T_X} L_{T_X} A_{T_X}^T v = \lambda A_{T_X} L_{T_X}^{1/2} L_{T_X}^{1/2} A_{T_X}^T v,$$

then multiplying by $(A_{T_X} L_{T_X}^{1/2})^{-1}$ from the left and by $\mathbf{I}$ in the middle we transform it to

$$L_{T_X}^{-1/2} A_{T_X}^T B_X A_{T_X} L_{T_X}^{-1/2} L_{T_X}^{1/2} A_{T_X}^T v = \lambda L_{T_X}^{1/2} A_{T_X}^T v.$$

Now, substituting

$$\tilde{B} = L_{T_X}^{-1/2} A_{T_X}^T B_X A_{T_X} L_{T_X}^{-1/2} = \left(L_{T_X}^{-1/2} A_{T_X}^T\right) B_X \left(L_{T_X}^{-1/2} A_{T_X}^T\right)^T$$

and

$$\tilde{v} = L_{T_X}^{1/2} A_{T_X}^T v \tag{16}$$

we get a standard eigenproblem for $\tilde{B}$

$$\tilde{B}\tilde{v} = \lambda\tilde{v}. \tag{17}$$

Solving (17) and using the inverse transformation of (16)

$$v = A_{T_X} L_{T_X}^{-1/2} \tilde{v}, \tag{18}$$

we obtain the solution v to the original problem (14), corresponding to the same eigenvalue λ. In particular, it proves that with our model assumptions (14) can be reduced to a standard eigenproblem

$$T_X^{-1} B_X v = \lambda v, \tag{19}$$

which takes the matrix form of

$$\left(T_X^{-1} B_X\right) V = V L, \tag{20}$$

where $L \in \mathbb{R}^{d\times d}$ is a diagonal matrix of eigenvalues in a non-decreasing order and $V \in \mathbb{R}^{d\times d}$ is a matrix of their corresponding column eigenvectors.

Note that there is another alternative formulation of the problem (14) via canonical correlation analysis (CCA), which may also come as a convenient way to see the task. In this setup Fisher's eigenvalues correspond to squared canonical correlation coefficients. We will not describe it here in details but we give references for interested readers. The approach, referred to as canonical discriminant analysis (CDA), was first mentioned in [21] and thoroughly described in [22]. The overview of classical CCA is given for instance in [12].

3.3. **Motivation.** What we refer to as Fisher's distinctness measure was inspired by [23], where the idea of using the eigenproblem formulation of the Fisher's discrimination task and its respective eigenvalues for assessing certain properties of data was used.

As explained in Subsection 3.2, Fisher's discriminant task can be stated in terms of eigenproblem given by (20). Then, its $(k-1)$ eigenvectors corresponding to the $(k-1)$ non-zero eigenvalues span the Fisher's subspace. Note that there are $k-1$ non-zero eigenvalues as according to the model assumptions $\operatorname{rank}(T_X) = d$ and $\operatorname{rank}(B_X) = k-1$ and $d > k-1$. Due to (20) we have

$$V^T T_X^{-1} B_X V = L,$$

so the eigenvalues capture variability in the spanning directions. As Fisher's task is scale invariant, the increase in variability may only be due to increase in between cluster scatter or decrease in within cluster scatter so it is expected to capture increase in structure distinctness very well. As squared canonical correlation coefficients (see references in Subsection 3.2), the eigenvalues remain in the interval of $[0, 1]$ which also makes them easy to compare and interpret. Additionally, except for being easy to compute numerically, they are also convenient to handle analytically, so they can easily be used in simulations as well as formal derivations and justifications. What remains, is to propose function of the eigenvalues that could serve as structure distinctness coefficient and analyze its performance. This was done by means of simulation study and described in the next section.

4. Simulation study

4.1. **Overview.** Due to its analytical complexity (1) is virtually intractable for mixtures with varied covariance matrices (heterogeneous) or of higher space dimension. However, it relatively easy undergoes simulations of Monte Carlo kind and can easily be approximated numerically with the best linear approximation described in subsection 2.2. As such, it may be used as a reference measure and replaced with another coefficient that reflects its behavior but offers the advantage of being computable and analytically tractable, also in a more complex setup.

The study was divided into two parts. In the first part two dimensional case was studied in details. Normal distribution was parametrized in a way that allowed for easy parameter control. Then all the possible combinations were tested and the influence of change in between cluster separation and within cluster dispersion was analysed. Three possible structure distinctness measures were compared — exact integral measure (1), its best linear approximation described in subsection 2.2 and Fisher's eigenvalue. For two dimensional data, the maximum number of two clusters was analysed (due to the assumption of $d > k-1$), which led to one dimensional projections. Therefore, there was just single Fisher's eigenvalue to compare so the two dimensional step could not give grounds for function selection. The two dimensional study served as a thorough assessment of single Fisher's eigenvalue performance.

In the second step multidimensional data was analyzed. Due to high number of possible mixture parameter combinations only a random selection was considered. This step was meant to confirm satisfactory performance of Fisher eigenvalues as input for structure distinctness measure. Higher dimensionality allowed for larger number of clusters, which resulted in $(k-1) > 1$ dimensionality of Fisher's subspace. As such, it also gave grounds for selecting appropriate function to transform $(k-1)$ eigenvalues into a single structure distinctness coefficient. Minimum $\lambda^X_{\min}$ and average $\bar{\lambda}^X$ over Fisher's non-zero eigenvalues were calculated as follows

$$\lambda^X_{\min} = \min_{j\in\{1,\dots,k-1\}} \lambda_j^{T_X^{-1}B_X} \tag{21}$$

and

$$\bar{\lambda}^X = \frac{1}{k-1}\sum_{j=1}^{k-1} \lambda_j^{T_X^{-1}B_X}, \tag{22}$$

and compared with the Monte Carlo estimates of the integral measure (1). Note that due to the larger number of classes allowed, wider comparisons with the best linear separator, defined for $k = 2$ only, were infeasible.

Note that although the original concept (1) is defined in terms of overlap (similarity) between the components, what is naturally captured by either minimum or average over non-zero Fisher's eigenvalues, reflects the opposite behavior, so should rather be referred to as distinctness (dissimilarity) measure. Therefore we compare it with $(1 - \mathrm{MLE}_{\mathrm{err}})$ (or $(1 - \mathbb{P}_{\mathrm{minmax}})$), which is the probability of correct MLE classification (or its best linear approximation). The transition from one to another is typically straightforward, however we point that out explicitly to avoid confusion or additional transformations of the coefficients.

Algorithm 4.1: TwoDimensionalDataGeneration$(r, \alpha, \lambda, q, k, N[])$

```
for each cluster $l \in \{1, \dots, k\}$
do
    comment: Determine cluster center $\mu$
    $\mu \leftarrow \left(r \cdot \sin\left((l-1)\cdot\frac{2\pi}{k}\right),\ r \cdot \cos\left((l-1)\cdot\frac{2\pi}{k}\right)\right)$
    comment: Compute covariance matrix $\mathbf{\Sigma}$
    $D \leftarrow \mathrm{diag}(\lambda, q\cdot\lambda)$   comment: dispersion and shape matrix
    $R \leftarrow \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix}$   comment: rotation matrix
    $\mathbf{\Sigma} \leftarrow RDR^T$
    comment: Generate data
    draw $N[l]$ observations
    add cluster mean $\mu$ to each observation
return ($data$)
```

4.2. **Two-dimensional simulations.** To allow for easy control over mixture parameters, two dimensional mixture density was parametrized in a convenient way. Cluster centers were located on a circle around origin $(0, 0)$ with radius r that controlled between cluster distance. To allow for heterogeneity, for each cluster covariance matrix was determined separately. Within cluster dispersion was captured by the leading eigenvalue $\lambda = \lambda_1$, cluster shape by eigenvalues' ratio $q = \lambda_2/\lambda_1$, and cluster rotation by rotation angle α. Based on these parameters for each component mean vector and covariance matrix were computed. For each component the data was generated with the algorithm based on Cholesky decomposition, using affine transformation property for multivariate normal distribution. The detailed description of the algorithm is provided in [24]. Assuming the number of clusters is given by k and $N \in \mathbf{R}^k$ contains desired cluster sizes, the above algorithm presents subsequent steps of data generation.

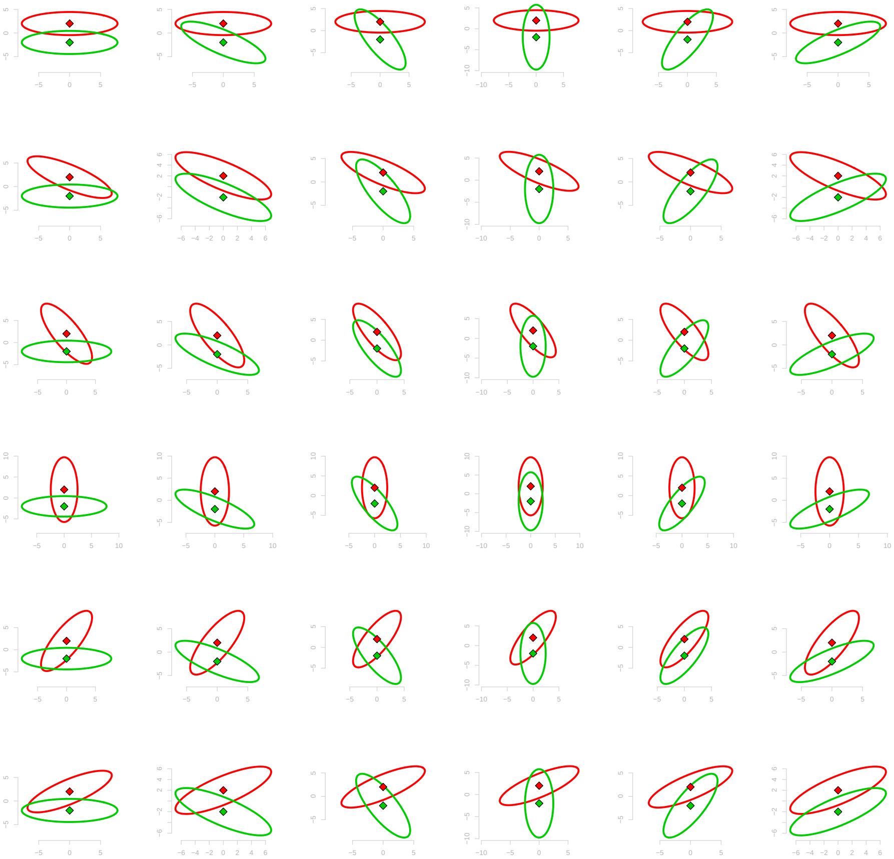

FIGURE 4. Design of two dimensional simulations — components' position with respect to each other.

The simulation design is shown in Figure 4, which presents all possible combinations of component position with respect to each other. Each of $i = 1, \ldots, 6$ rows corresponds to $i \cdot \pi/6$ angle rotation for the first (red) component, while each of $j = 1, \ldots, 6$ columns corresponds to $j \cdot \pi/6$ angle rotation for the second (green) component. Altogether it yields 36 basic mixture positions. For each position an influence of a single factor is analyzed and this includes in particular – increase in between cluster distance (Figures 7 and 8), increase in within cluster dispersion for both (Figures 9 and 10) and for first (Figures 11 and 12) and second (Figures 13 and 14) spanning direction only. The special case of spherical clusters is analyzed separately (Figures 15 to 18). All the results are available in Appendix, Section A.

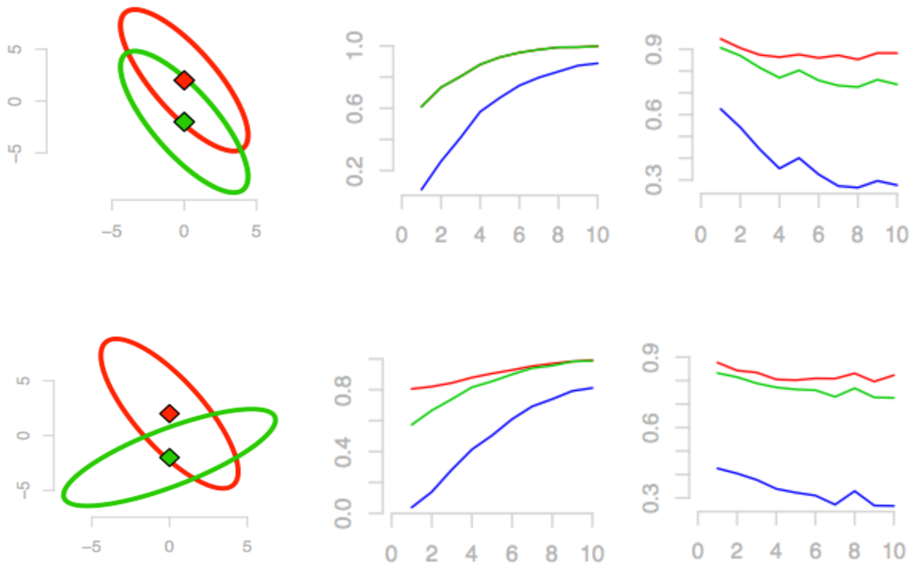

Figure 5. Impact of increasing between cluster distance (second column) and within cluster dispersion (third column) for mixtures in positions as indicated in the first column. Red line indicates exact (integral) structure distinctness, green — its linear approximate and blue — Fisher's eigenvalue.

Example of what can be observed in all the charts is shown in Figure 5. Even though the values for Fisher's eigenvalue are much smaller, their variability reflects behavior of the integral measure to a large extent. It is even more in tune with the linear estimate, which is to be expected given the linear nature of the Fisher's discrimination task. Note, that the best linear approximate gives the upper bound on the precision with which a linear concept may reflect behavior of the non-linear integral measure. Also, it gives upper limit on classification accuracy using linear classifiers, which is the case of Fisher discriminant. Note also, that the component position in the upper row indicates homogeneity (i.e. equal covariance matrices for both components). This property is lost when within cluster variability increases for one of the components (last column). However, it remains when only between cluster distance is affected (middle column). Therefore, exact integral measure and its linear estimate overlap in this case.

4.3. Multi-dimensional simulations. In higher dimensions direct analytical control over distance and dispersion of mixture parameters is much more complex. Additionally, there are many more combinations to examine. As such, the simulations were reduced to randomly chosen mixture parameters' combinations corresponding to the mixture position. For each position the impact of increasing between cluster distance and within cluster dispersion was analysed. The study was designed to verify adequacy of the information carried by the Fisher's eigenvalues and to select its appropriate function to serve as the structure distinctness coefficient. Results are attached in Appendix A in Figures 19 to 22. In each row charts for random but fixed set of cluster means are presented. Similarly, the set of covariance matrices is random but fixed in each column. Mean vectors and covariance matrices in d dimensions were determined using **R** package **clusterGeneration**, which implements the ideas described in [25] and [26]. Additionally, mean coordinates are re-scaled to lie in the interval $[-3\sqrt{d}, 3\sqrt{d}]$ which corresponds to the range of the maximum three standard deviations for covariance matrix. As such, the possible overlap between components stretches from complete to negligible.

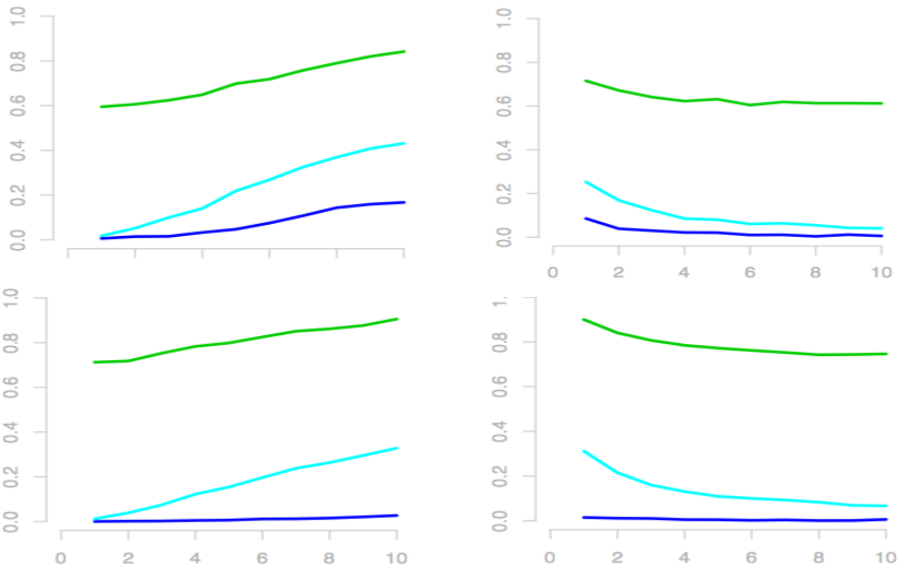

FIGURE 6. Effect of increasing between cluster distance (left column) and within cluster dispersion (right column). Upper row gives results for three dimensional simulations, while bottom row for five dimensional case. Green line indicates Monte Carlo estimate of the integral structure distinctness, turquoise average non-zero Fisher's eigenvalue, while blue — Fisher's smallest non-zero eigenvalue.

Again, what can be observed in all the simulation plots in Appendix A is illustrated in Figure 6. Behavior of average Fisher's eigenvalue as given by (22) reflects variability of the integral measure. At the same time, minimum non-zero Fisher's eigenvalue (22) is less sensitive and therefore captures the changes in distinctness to a lesser extent, which becomes even more apparent as the number of dimensions increases. As such, the average non-zero Fisher's eigenvalue tends to outperform the minimum non-zero Fisher's eigenvalue and therefore the former shall be recommended as the distinctness coefficient.

5. Conclusions.

In this work we derive and motivate measure of distinctness (or alternatively – overlap) between clusters of data, generated from a Gaussian mixture model. The approach uses alternative formulation of Fisher's discrimination task, which is stated in terms of a generalized eigenproblem. We show the task is well posed in the context of the assumed model and can be reduced to a standard eigenproblem with real eigenvalues. We then express the distinctness coefficient as the average eigenvalue over the non-zero eigenvalues of the solution. We compare the behavior of the coefficient with the generic (integral) measure of structure distinctness defined in terms of the actual overlap between the corresponding distributions and its best linear approximation. Although the values of the Fisher's coefficient are lower than the values of actual overlap, their dynamic reflects very well the behavior of the generic integral measure and even better – its best linear approximation. As opposed to the generic integral measure and its best linear approximation, the Fisher's coefficient offers the advantage of being not only numerically easily computable but also analytically tractable, even in a complex setup, regardless of the dimensionality of the space and heterogeneity of covariance matrices.

References

[1] A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distributions, Bulletin of Cal. Math. Soc. 35 (1) (1943) 99–109.

[2] K. Fukunaga, Introduction to statistical pattern recognition, 2nd Edition, Computer Science and Scientific Computing, Academic Press, Inc., Boston, MA, 1990.

[3] N. E. Day, Estimating the components of a mixture of normal distributions, Biometrika 56 (3) (1969) 463–474.
URL http://www.jstor.org/stable/2334652

[4] G. J. McLachlan, K. E. Basford, Mixture models, Vol. 84 of Statistics: Textbooks and Monographs, Marcel Dekker, Inc., New York, 1988, inference and applications to clustering.

[5] S. Kullback, R. A. Leibler, On information and sufficiency, Ann. Math. Statistics 22 (1951) 79–86. doi:10.1214/aoms/1177729694.

[6] H. F. Inman, E. L. Bradley, Jr., The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities, Comm. Statist. Theory Methods 18 (10) (1989) 3851–3874. doi:10.1080/03610928908830127.
URL http://dx.doi.org/10.1080/03610928908830127

[7] S. Dasgupta, Learning mixtures of gaussians, in: 40th Annual Symposium on Foundations of Computer Science, 1999, pp. 634–644. doi:10.1109/SFFCS.1999.814639.

[8] R. Maitra, Initializing partition-optimization algorithms, IEEE/ACM Trans. Comput. Biol. Bioinformatics 6 (1) (2009) 144–157. doi:10.1109/TCBB.2007.70244.
URL http://dx.doi.org/10.1109/TCBB.2007.70244

[9] H.-J. Sun, M. Sun, S.-R. Wang, A measurement of overlap rate between gaussiancomponents, in: International Conference on Machine Learning and Cybernetics, Vol. 4, 2007, pp. 2373–2378. doi:10.1109/ICMLC.2007.4370542.

[10] S. Ray, B. G. Lindsay, The topography of multivariate normal mixtures, Ann. Statist. 33 (5) (2005) 2042–2065. doi:10.1214/009053605000000417.
URL http://dx.doi.org/10.1214/009053605000000417

[11] H. Sun, S. Wang, Measuring the component overlapping in the Gaussian mixture model, Data Min. Knowl. Discov. 23 (3) (2011) 479–502. doi:10.1007/s10618-011-0212-3.
URL http://dx.doi.org/10.1007/s10618-011-0212-3

[12] K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate analysis, Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York-Toronto, Ont., 1979, probability and Mathematical Statistics: A Series of Monographs and Textbooks.

[13] T. W. Anderson, R. R. Bahadur, Classification into two multivariate normal distributions with different covariance matrices, The Annals of Mathematical Statistics 33 (2) (1962) 420–431. doi:10.1214/aoms/1177704568.
URL http://dx.doi.org/10.1214/aoms/1177704568

[14] G. J. Székely, M. L. Rizzo, Hierarchical clustering via joint between-within distances: extending Ward's minimum variance method, J. Classification 22 (2) (2005) 151–183. doi:10.1007/s00357-005-0012-9.
URL http://dx.doi.org/10.1007/s00357-005-0012-9

[15] J. H. Ward, Jr., Hierarchical grouping to optimize an objective function, J. Amer. Statist. Assoc. 58 (1963) 236–244. doi:10.2307/2282967.

[16] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning, 2nd Edition, Springer Series in Statistics, Springer, New York, 2009, data mining, inference, and prediction. doi:10.1007/978-0-387-84858-7.
URL http://dx.doi.org/10.1007/978-0-387-84858-7

[17] S. Lipovetsky, Additive and multiplicative mixed normal distributions and finding cluster centers, International Journal of Machine Learning and Cybernetics 4 (1) (2013) 1–11. doi:10.1007/s13042-012-0070-3.
URL http://dx.doi.org/10.1007/s13042-012-0070-3

[18] S. Lipovetsky, Total odds and other objectives for clustering via multinomial-logit model, Advances in Adaptive Data Analysis 04 (03) (2012) 1250019. doi:10.1142/S1793536912500197.
URL http://www.worldscientific.com/doi/abs/10.1142/S1793536912500197

[19] S. Lipovetsky, Finding cluster centers and sizes via multinomial parameterization, Applied Mathematics and Computation 221 (2013) 571–580. doi:10.1016/j.amc.2013.06.098.

[20] R. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (2) (1936) 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.
URL http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x

[21] M. S. Bartlett, Further aspects of the theory of multiple regression, Mathematical Proceedings of the Cambridge Philosophical Society 34 (1938) 33–40. doi:10.1017/S0305004100019897.
URL http://journals.cambridge.org/article_S0305004100019897

[22] W. Dillon, M. Goldstein, Multivariate analysis: methods and applications, Wiley series in probability and mathematical statistics: Applied probability and statistics, Wiley, 1984.

[23] S. Brubaker, S. Vempala, Isotropic pca and affine-invariant clustering, in: M. Grötschel, G. Katona, G. Sági (Eds.), Building Bridges, Vol. 19 of Bolyai Society Mathematical Studies, Springer Berlin Heidelberg, 2008, pp. 241–281. doi:10.1007/978-3-540-85221-6_8.
URL http://dx.doi.org/10.1007/978-3-540-85221-6_8

[24] J. E. Gentle, Random number generation and Monte Carlo methods, 2nd Edition, Statistics and Computing, Springer, New York, 2003.

[25] H. Joe, Generating random correlation matrices based on partial correlations, J. Multivariate Anal. 97 (10) (2006) 2177–2189. doi:10.1016/j.jmva.2005.05.010.
URL http://dx.doi.org/10.1016/j.jmva.2005.05.010

[26] D. Kurowicka, R. Cooke, Uncertainty analysis with high dimensional dependence modelling, Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd., Chichester, 2006. doi:10.1002/0470863072.
URL http://dx.doi.org/10.1002/0470863072

Appendix A. Simulation results

Ewa Nowakowska, Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
E-mail address: `ewa.nowakowska@ipipan.waw.pl`

Jacek Koronacki, Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
E-mail address: `jacek.koronacki@ipipan.waw.pl`

Stan Lipovetsky, GfK Custom Research North America, Marketing & Data Sciences, 8401 Golden Valley Rd., Minneapolis MN 55427, USA
E-mail address: `stan.lipovetsky@gfk.com`

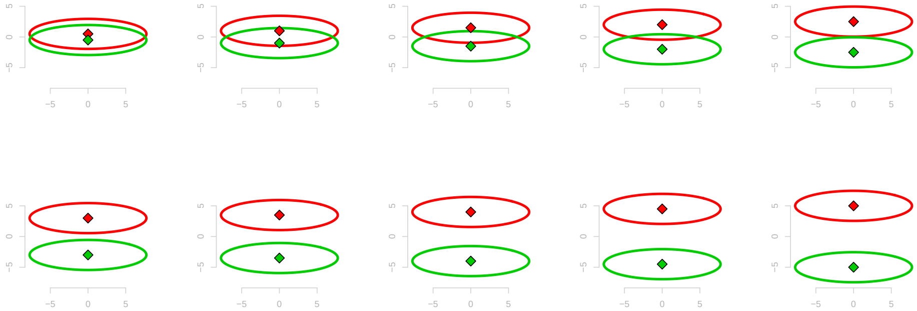

FIGURE 7. Diagram of increasing between cluster **distance**

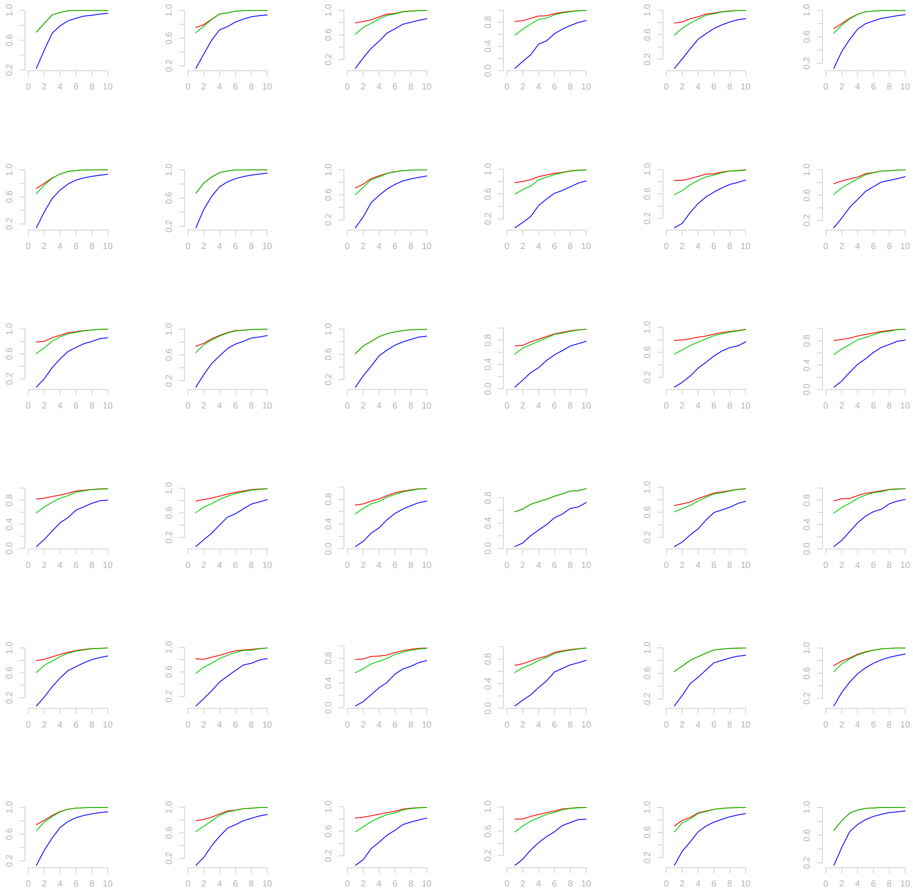

FIGURE 8. For clusters in position as in Figure 4, each chart presents impact of increasing between cluster distance according to the pattern from Figure 7, measured with exact $(1 - \mathrm{MLE}_{\mathrm{err}})$ (red), its linear approximation $(1 - \mathbb{P}_{\mathrm{minmax}})$ (green) and Fisher's eigenvalue (blue).

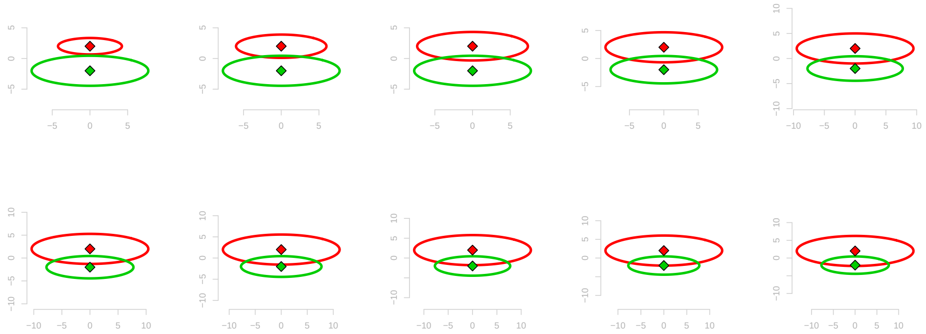

FIGURE 9. Diagram of increasing within cluster **dispersion – in both spanning directions**

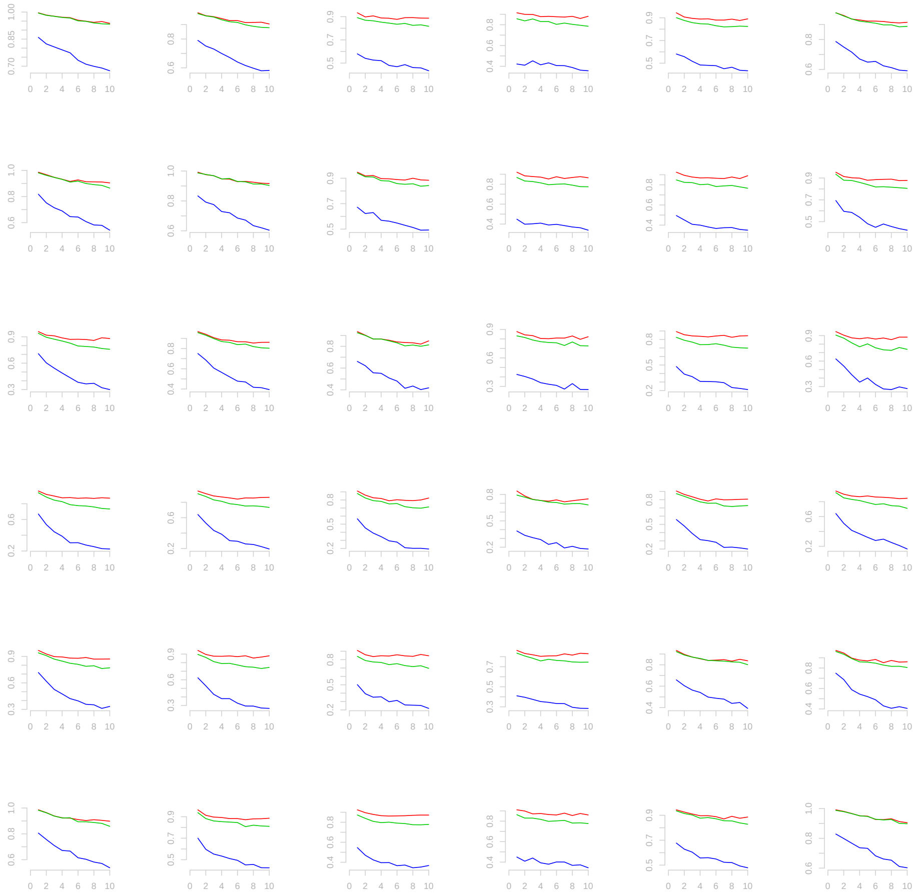

FIGURE 10. For clusters in position as in Figure 4, each chart presents impact of increasing within cluster dispersion (both directions) according to the pattern from Figure 9, measured with exact $(1 - \mathrm{MLE}_{\mathrm{err}})$ (red), its linear approximation $(1 - \mathbb{P}_{\mathrm{minmax}})$ (green) and Fisher's eigenvalue (blue).

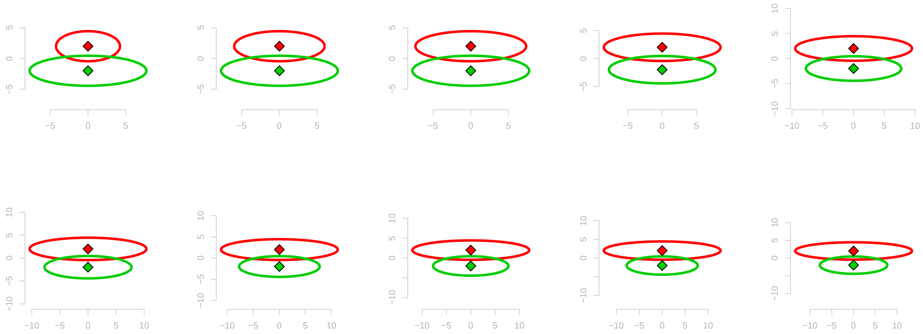

FIGURE 11. Diagram of increasing within cluster **dispersion** – **first spanning direction**

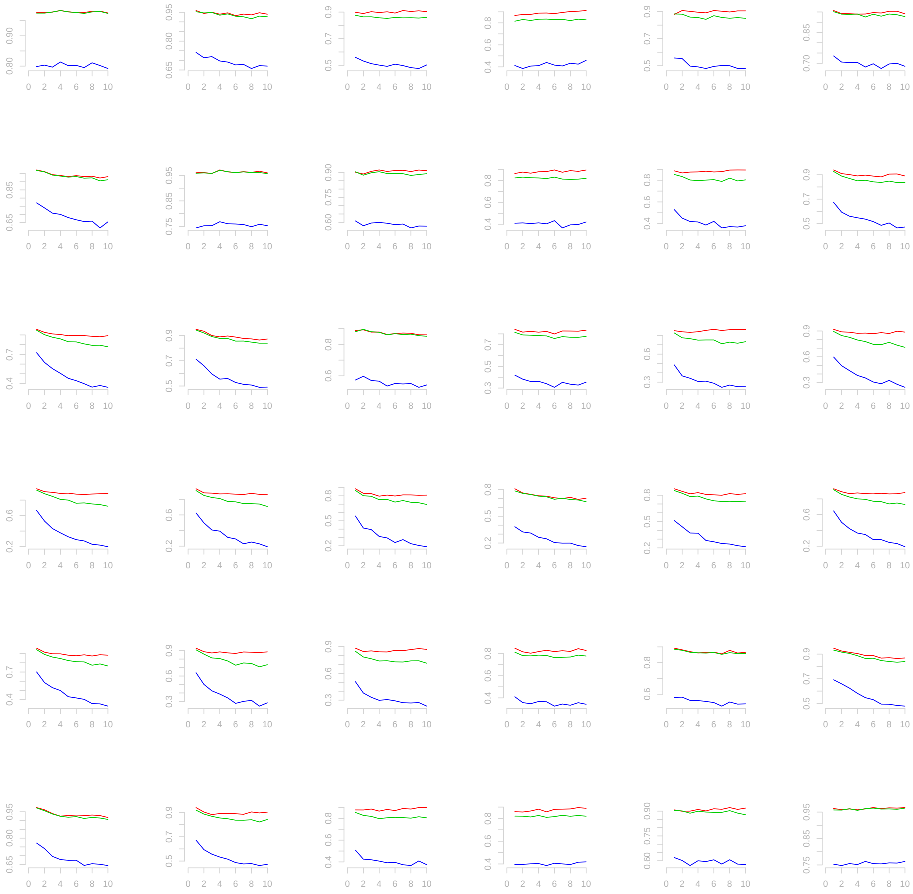

FIGURE 12. For clusters in position as in Figure 4, each chart presents impact of increasing within cluster dispersion (first direction) according to the pattern from Figure 11, measured with exact $(1 - \text{MLE}_{\text{err}})$ (red), its linear approximation $(1 - \mathbb{P}_{\text{minmax}})$ (green) and Fisher's eigenvalue (blue).

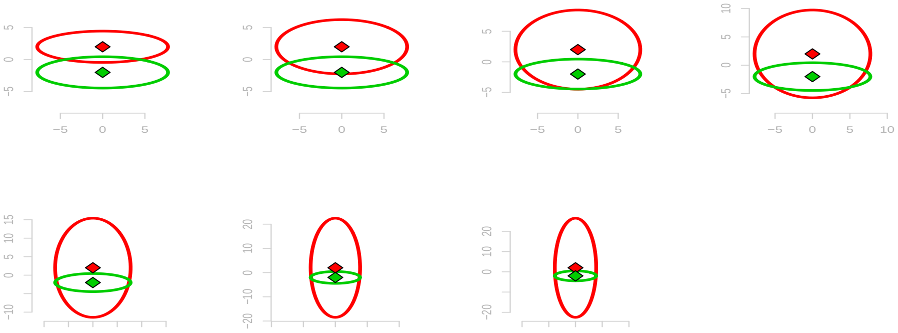

FIGURE 13. Diagram of increasing within cluster **dispersion – second spanning direction**

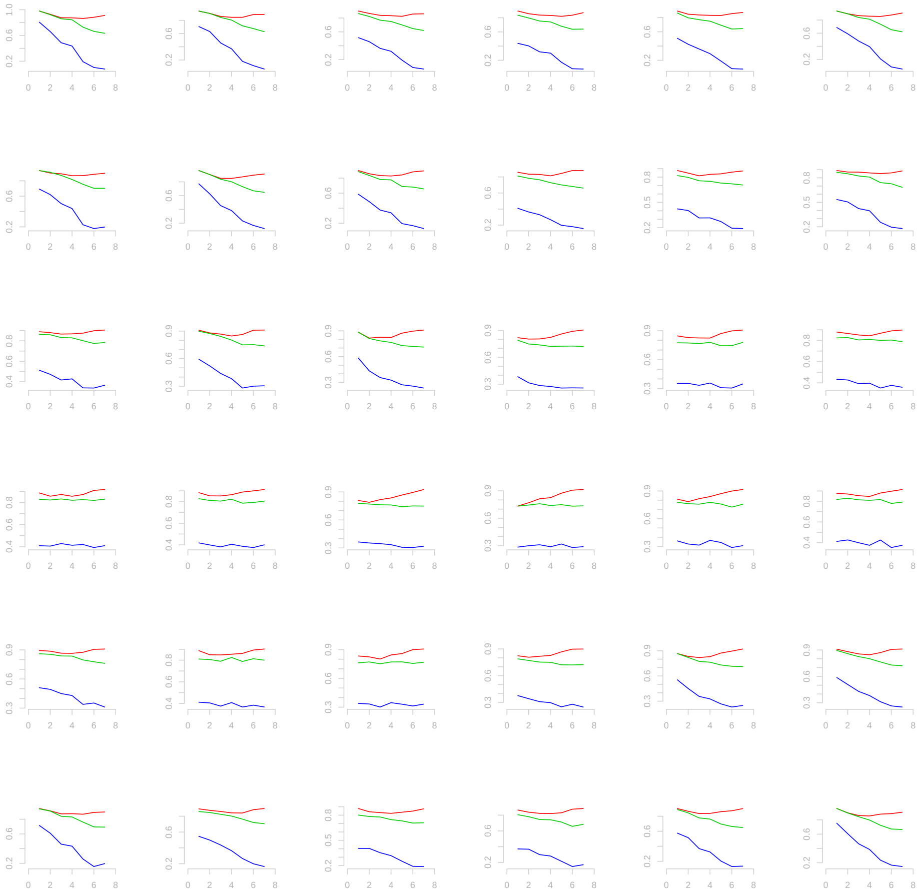

FIGURE 14. For clusters in position as in Figure 4, each chart presents impact of increasing within cluster dispersion (second direction) according to the pattern from Figure 13, measured with exact $(1 - \mathrm{MLE}_{\mathrm{err}})$ (red), its linear approximation $(1 - \mathbb{P}_{\mathrm{minmax}})$ (green) and Fisher's eigenvalue (blue).

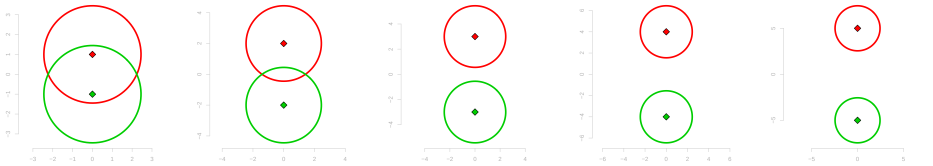

FIGURE 15. **Spherical components** – diagram of increasing between cluster **distance**

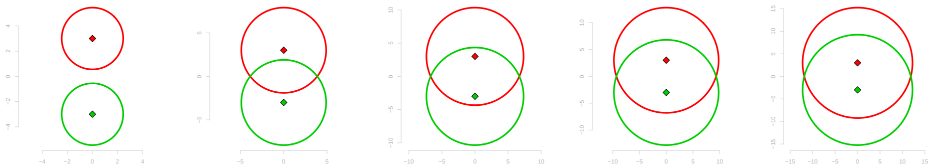

FIGURE 16. Diagram of **balanced** increase in within cluster **dispersion** (same increase for both clusters)

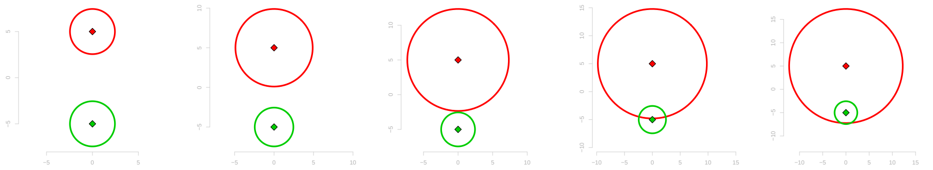

FIGURE 17. Diagram of **unbalanced** increase in within cluster **dispersion** (increase for one cluster only)

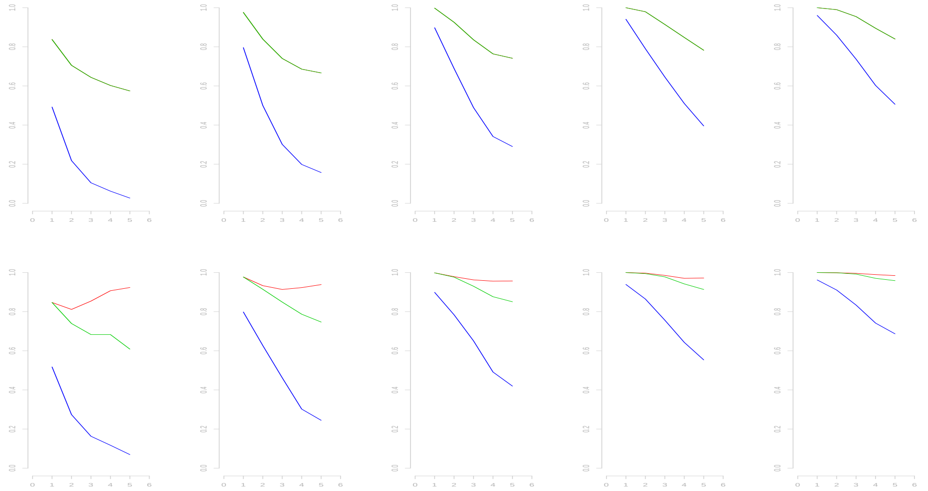

FIGURE 18. For spherical clusters, each line follows the distance pattern from Figure 15, each chart presents impact of increasing within cluster dispersion – balanced in the first line (Figure 16, unbalanced in the second (Figure 17), measured with exact $(1 - \mathrm{MLE}_{\mathrm{err}})$ (red), its linear approximation $(1 - \mathbb{P}_{\mathrm{minmax}})$ (green) and Fisher's eigenvalue (blue).

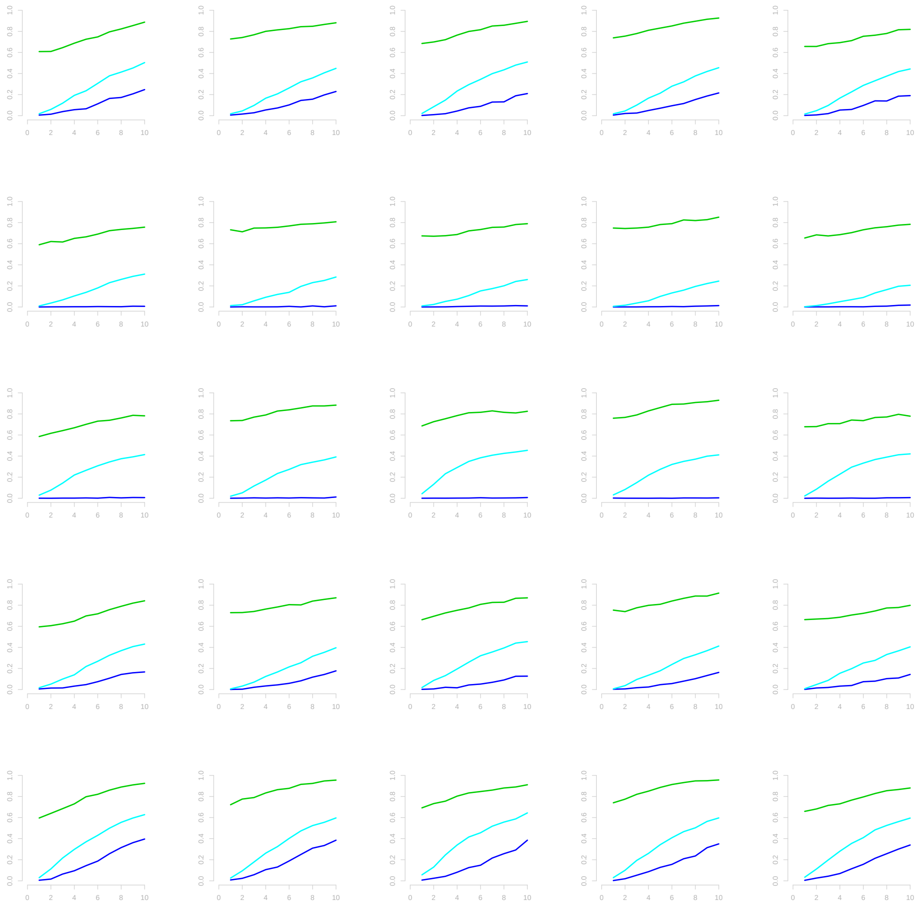

FIGURE 19. Three dimensions: for random (but fixed in each row) set of cluster means and random (but fixed in each column) set of covariance matrices, each chart presents impact of increasing between cluster **distance**, measured with exact integral measure (green), average Fisher's eigenvalue (turquoise) and minimum Fisher's eigenvalue (blue).

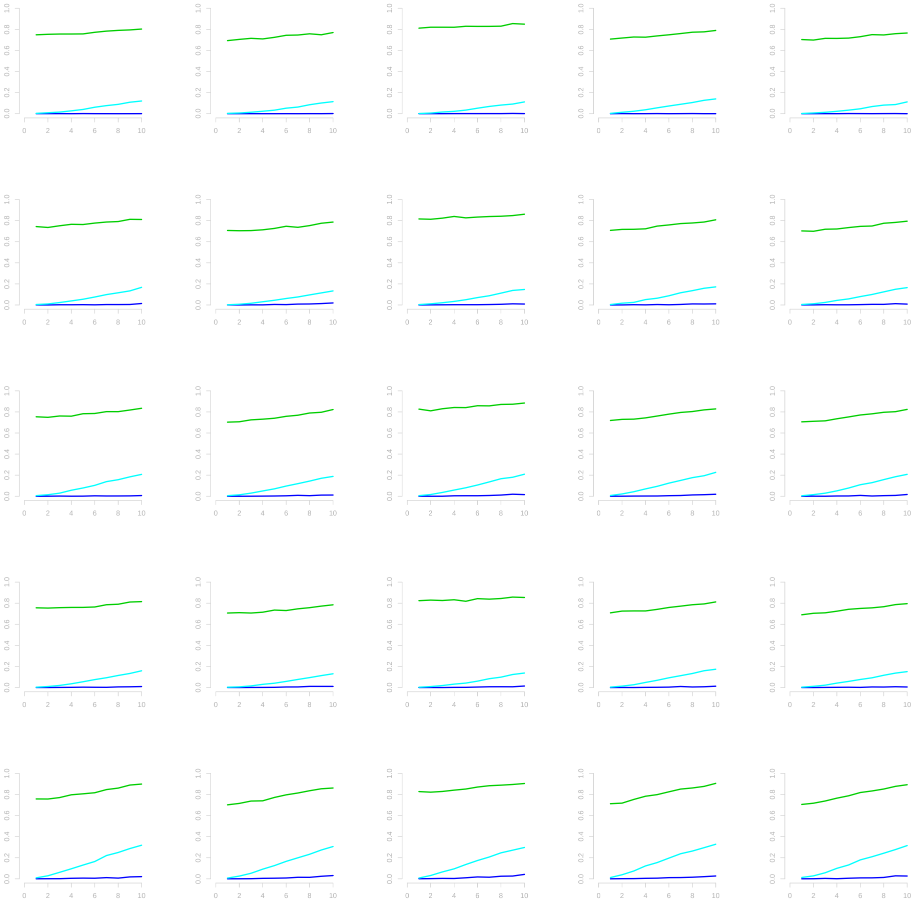

Figure 20. Five dimensions: for random (but fixed in each row) set of cluster means and random (but fixed in each column) set of covariance matrices, each chart presents impact of increasing between cluster **distance**, measured with exact integral measure (green), average Fisher's eigenvalue (turquoise) and minimum Fisher's eigenvalue (blue).

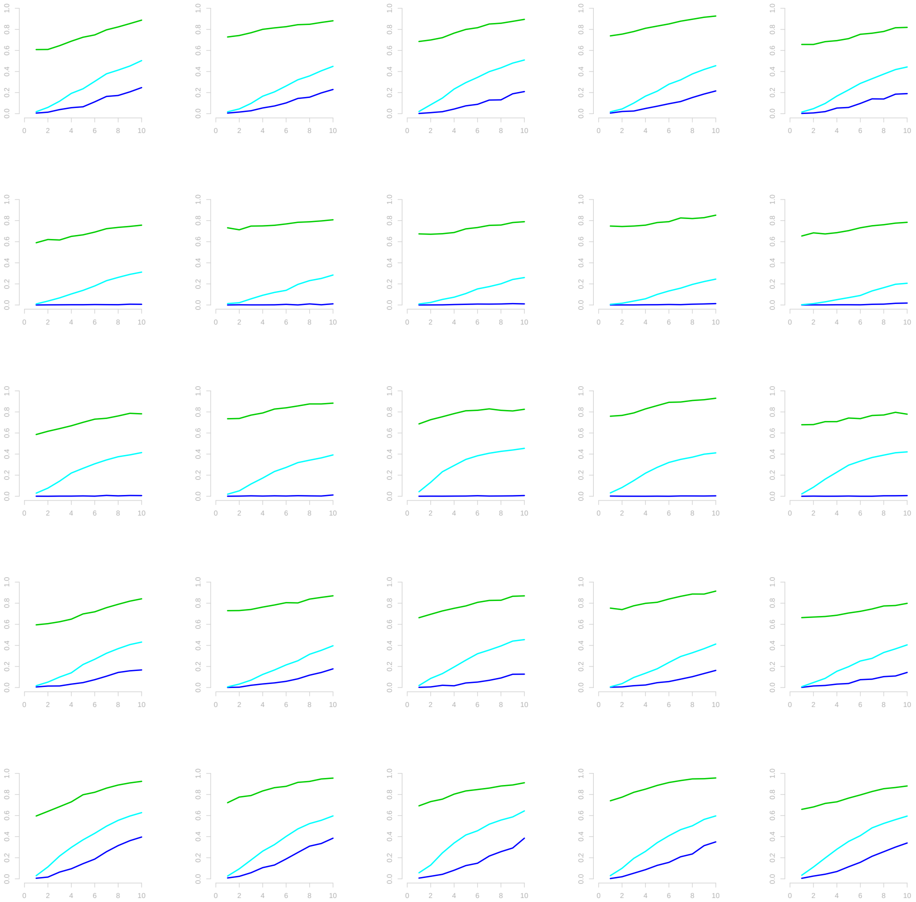

FIGURE 21. Three dimensions: for random (but fixed in each row) set of cluster means and random (but fixed in each column) set of covariance matrices, each chart presents impact of increasing within cluster **dispersion**, measured with exact integral measure (green), average Fisher's eigenvalue (turquoise) and minimum Fisher's eigenvalue (blue).

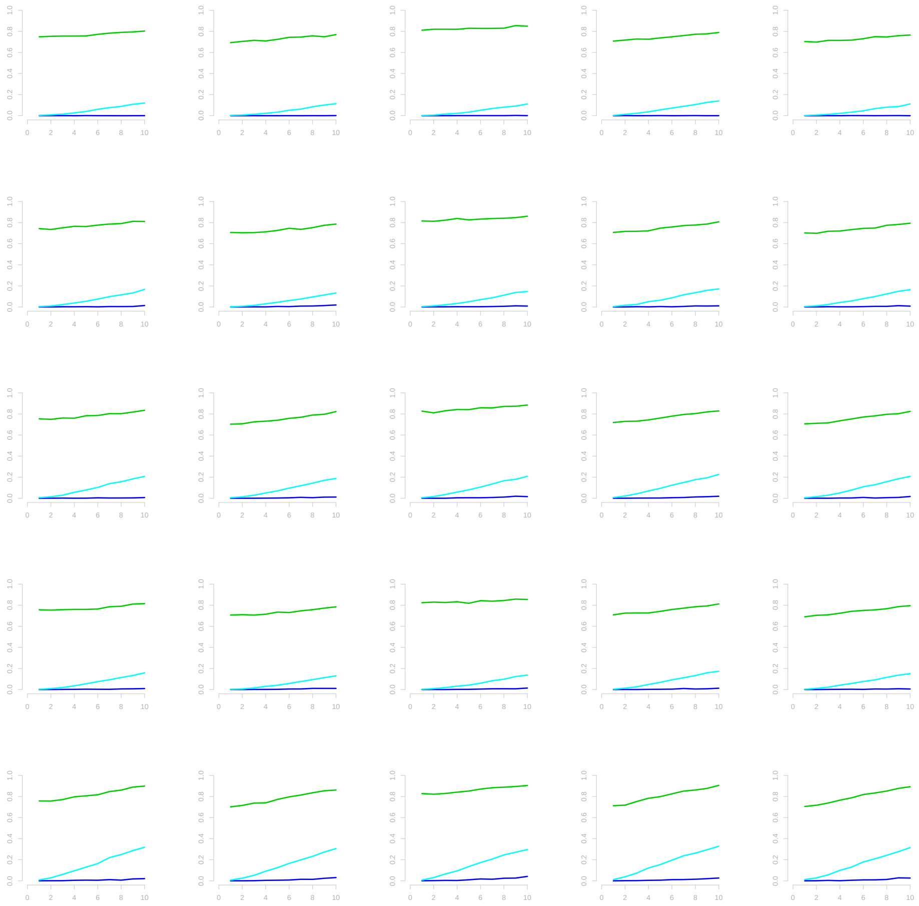

Figure 22. Five dimensions: for random (but fixed in each row) set of cluster means and random (but fixed in each column) set of covariance matrices, each chart presents impact of increasing within cluster **dispersion**, measured with exact integral measure (green), average Fisher's eigenvalue (turquoise) and minimum Fisher's eigenvalue (blue).