BAYESIAN GROUP ANALYSIS

W. VON DER LINDEN, V. DOSE AND A. RAMASWAMI
Max-Planck-Institut f"ur Plasmaphysik, EURATOM Association
D-85740 Garching b. M"unchen, Germany
e-mail: wvl@ipp.mpg.de

Abstract.
In many fields of research the following problem is encountered: a large collection of data is given for which a detailed theory is yet missing. To gain insight into the underlying problem it is important to reveal the interrelationships in the data and to determine the relevant input and response quantities. A central part of this task is to find the natural splitting of the data into groups and to analyze the respective characteristics. Bayesian probability theory is invoked for a consistent treatment of these problems. Due to Ockham's Razor, which is an integral part of the theory, the simplest group configuration that still fits the data has the highest probability. In addition the Bayesian approach allows to eliminate outliers, which otherwise could lead to erroneous conclusions. Simple textbook and mock data sets are analyzed in order to assess the Bayesian approach.

Key words: Auto-classification,auto-clustering,group analysis,Mahalonobis distance

1. Introduction

The type of problems we are addressing in this paper can best be explained by the simple computer vision problem depicted in fig.1. There is a collection of data points characterized by certain characteristics, here x- and y-coordinate. Our goal is to collect the data into groups, characterized by straight lines. Data in a group will be statistically distributed around a perfect straight line due to natural scatter and experimental uncertainties. A couple of questions arises immediately: How many groups (straight lines) are there? What are their characteristics, i.e. intercept and slope? Which data point belongs to which group? Are there outliers, i.e. individual points that don't fit into any of the other groups? This simple example reveals already a severe shortcoming of maximum-likelihood approaches: the perfect fit would consist of two-element groups. Obviously, this solution is far from what intuition tells us when considering fig.1. Moreover, this solution is completely useless, since we don't learn anything about the underlying problem and it is downright wrong since we know that there must be a scatter about the intrinsic straight lines. So a crucial element is missing to counteract the data-constraints: the tendency towards simplicity. In Bayesian probability theory this tendency is automatically taken care of as a result of the prior probabilities in parameter space. The gain in likelihood volume is balanced by the loss of prior volume. Why carry 100 baskets if there is only one egg to fit in?

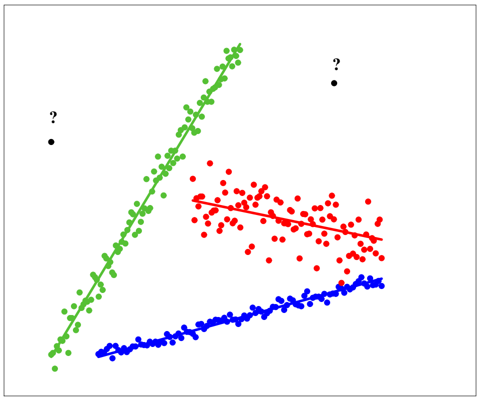

Figure 1. Simple computer vision problem for illustration.

Now let us make the problem more abstract but at the same time more realistic. Very often in science, especially in the beginning of a new field when little or nothing is known and a detailed theory is missing, data are gathered in the hope to gain insight into the underlying problem. Specifically one would like to unravel the interrelationships present in the data and to identify the relevant input and response quantities. Part of the learning process is to find the natural groupings in the data. A great deal can be learned from those characteristics that discriminate the groups.

Many – more or less ad-hoc – methods can be found in the literature for the type of problems we are discussing (see e.g. [1]). In this paper we tackle these problems by strictly following the rules of probability theory and without resorting to any ad-hoc methods or free parameters. The approach we use is a multivariate generalization of Larry Bretthorst's study *On the difference in means* [2]. In the probabilistic approach we will find two competing factors. The likelihood term, which quantifies the data-constraints, would try to find the least deviation from the model. As discussed above, this extreme case is of little use and has no predictive power. The introduction of the prior automatically enforces a tradeoff between the fit of the data and the complexity of the group, giving an automatic form of Ockham's Razor. The latter is inherent in Bayesian probability theory and favors a small number of groups.

For many scientific applications it is, however, not desirable to analyze blindly all combinatorially possible groupings to say nothing of the gigantic complexity.

A very promising approximate evaluation of the auto-classification problem using bayesian probability theory has been suggested recently [3] and implemented in a program package called AutoClass. Here we are aiming at the following type of problems: a set of data points is taken under the same experimental conditions. The external, non-scalable parameters are only changed for different sets. Due to the lack of a detailed theory it is, however, not clear from the outset, which of these parameters or which combination of changes really influences the features of the underlying problem. On the contrary, these are precisely the points one wishes to unravel. Therefore, the question to be answered is which of these data sets are different and which can be pooled into one group in order to learn about the relevant input and response quantities and there interrelationships.

For a quantitative analysis it is also necessary to automatically identify outliers in order to avoid over-rating their influence in the analysis of group characteristics. For a quantitative study, we have to specify the mathematical meaning of a group. We are not ambitious enough to head for a theory of everything and restrict our theory to linear problems. I.e. we consider groups which can be described by hyperplains in the space of the data-characteristics, like in fig.1. The restriction is actually not as severe as it looks at first glance. Many problems are globally linear, many others are locally linear in the regime for which data points are available, and many apparently nonlinear problems can be made linear by a suitable transformation. An object is characterized by m characteristics $\mathbf{d} = (d^1, \ldots, d^m)$ which are represented by real numbers, like height, length and mass. We consider only continuous characteristics. The discrete characteristics are naturally used to preselect groups. A group, ν say, is characterized – according to the above assumption – by mean vector $\mathbf{m}_\nu$ and covariance matrix C_ν. The covariance originates from the natural scatter and measurement errors. The data points $\mathbf{d}^i_\nu$ (# i in group ν) are i.i.d. multi-normal:

$$p(\mathbf{d}^i_\nu|C_\nu, \mathbf{m}_\nu, \mathcal{I}) \quad = \quad |2\pi C_\nu|^{-1/2} \mathrm{etr}\left\{-\frac{1}{2}C_\nu^{-1}(\mathbf{d}^i_\nu - \mathbf{m}_\nu)(\mathbf{d}^i_\nu - \mathbf{m}_\nu)^T\right\} \qquad ,$$

with the definition $\mathrm{etr}\,\{X\} = \exp(\mathrm{tr}(X))$. By virtue of the conditional independence the likelihood for all data-points in group ν reads

$$\begin{aligned} p(\{\mathbf{d}_\nu\}|C_\nu, \mathbf{m}_\nu, \mathcal{I}) &= |2\pi C_\nu|^{-n/2} \mathrm{etr}\left\{-\frac{1}{2}C_\nu^{-1}\overline{C_\nu}\right\} \\ \overline{C_\nu} &= \sum_{i=1}^{n}(\mathbf{d}^i_\nu - \mathbf{m}_\nu)(\mathbf{d}^i_\nu - \mathbf{m}_\nu)^T \end{aligned}$$

$\overline{C_\nu}$ is proportional to the sample covariance. The different notation has been used to simplify the expressions.

2. Bayesian Approach

Now that the problem is uniquely specified we head for the Bayesian approach. The questions raised in the introduction translate into hypotheses H_n which contain the following information

- the data belong to n different groups as specified by d^i_ν
- the number of elements in group ν is N_ν.
- group ν has mean $\mathbf{m}_\nu$ and covariance C_ν
- the means $\mathbf{m}_\nu$ of different groups are different and unknown i.e. they are nuisance parameters bound to be marginalized over.
- further specification of hypothesis H_n may consider the cases
  1. covariance matrices are the same / different in all groups
  2. covariance matrices are known / unknown.

The probability for hypothesis H_n can be determined via Bayes' theorem:

$$P(H_n|D,\mathcal{I}) \propto p(D|H_n,\mathcal{I})\ P(H_n|\mathcal{I}) \qquad . \tag{1}$$

The desired probability factors into marginal likelihood $p(D|H_n,\mathcal{I})$ and prior $P(H_n|\mathcal{I})$. The latter cannot be simplified further and has to be assigned according to the problem under consideration. The proportionality constant drops out in model comparison and needs no further discussion.

2.1. MARGINAL LIKELIHOOD $P(D|H_N,\mathcal{I})$

We assume for the moment that the covariance matrices are known. In this case the marginal likelihood entering Eq.1 is $p(D|H_n,\{C_\nu\},\mathcal{I})$. One bit of information is, however, still missing to make the likelihood complete, namely the true means. They are introduced via the marginalization rule:

$$p(D|H_n,\{C_\nu\},\mathcal{I}) \quad = \quad p(\{\mathbf{m}_\nu^{\mathrm{ML}}\}|H_n,\mathcal{I}) \prod_{\nu=1}^{n} \int d^L\mathbf{m}_\nu p(D_\nu|H_n,C_\nu,\mathbf{m}_\nu,\mathcal{I}) \tag{2}$$

D_ν are all data points in group ν. Several remarks are in order to explain Eq.2. We employed the approach first introduced by Larry Bretthorst for the bayesian model selection [2,4–6]. Although the mathematical route he takes is different the idea behind it is identical to what we discuss next. We assumed, and this is actually part of our background information $\mathcal{I}$, that the prior $p(\mathbf{m}_\nu|H_n,\mathcal{I})$ as function of $\mathbf{m}_\nu$ is slowly varying as compared to the likelihood $p(D_\nu|H_n,C_\nu,\mathbf{m}_\nu,\mathcal{I})$. The latter is dominated by a single peak, the maximum likelihood solution $\mathbf{m}_\nu^{\mathrm{ML}}$. If this is so, we can take the prior out of the integral like in Eq.2. In other words we consider only situations where the effort put into the experiment to measure the data pays off. As far as the parameters $\mathbf{m}_\nu$ are concerned, the prior has no influence in this approach but it is utterly important for model comparison since it provides the multiplicities for Occam's razor. Obviously for ill-posed problems the assumptions need some revision to allow for the small eigenvalues of the transformation from parameter to data space and to introduce a means for regularization. For the present type of problems, however, the assumptions are perfectly valid, otherwise the maximum-likelihood approaches used in orthodox statistics would fail badly. In the assumed situation of weak prior knowledge there is naturally a wide variety of functional forms conceivable and the result must not depend on the details of the prior. We construct the prior as outlined by Jaynes [7] and extensively used by Bretthorst[4–6,2] upon using testable information and the maximum entropy principle. Reasonable testable information in our case is the average value of the characteristics over all individuals irrespective of their group affiliation, i.e. $\overline{\mathbf{m}}$. Since the prior is conditional on H_n, the number of groups n as well as the number N_ν of points in group ν is known. We can write

$$\overline{\mathbf{m}} = \int p(\{\mathbf{m}_\nu\}|H_n, I) \left\{ \sum_\nu N_\nu \mathbf{m}_\nu \right\} d\mathbf{m}^{Ln} \qquad . \tag{3}$$

The maximum entropy principle [7] yields

$$p(\{\mathbf{m}_\nu\}|\lambda, I) = \prod_{i=1}^{L} \lambda_i^n \prod_{\nu=1}^{n} N_\nu^L \ \exp{(-\lambda^T \sum_{\nu=1}^{n} N_\nu \mathbf{m}_\nu)} \qquad , \tag{4}$$

where λ_i are Lagrange parameters to fulfill the constraint Eq.3. They are hyperparameters of the theory and are marginalized over to obtain

$$p(\{\mathbf{m}_\nu\}|H_n, I) \propto \left(\prod_{i=1}^{L} \frac{\sum_{\nu=1}^{n} N_\nu m_\nu^i}{(\Gamma(n) \prod_{\nu=1}^{n} N_\nu)^{1/n}} \right)^{-n} = \left(\prod_{i=1}^{L} L_i^{\mathrm{eff}} \right)^{-n} \qquad . \tag{5}$$

Jeffreys' prior has been used for the positive hyperparameters. The proportionality constant is independent of the number of groups n and the dimension L, respectively and deserves no further attention. The main purpose of the prior is to count the multiplicity. In this respect it can be cast into the form of an effective flat as indicated by the last equality in Eq.5 with an uncertainty in the i-th characteristics of L_i^{eff}.

2.2. POSTERIOR PROBABILITY FOR H_N IF $C_\nu = C$ IS KNOWN

We restrict the following discussion to the case that all groups have the same covariance matrix. A generalization to differing covariances for different groups is merely a matter of more indices and thorough book-keeping. If the common covariance matrix is even known then the derivation of the posterior probability for hypothesis H_n is completed

$$\begin{aligned} P(H_n|D, C, \mathcal{I}) \quad &\propto \quad P(H_n|\mathcal{I})\, P(\{\overline{d}_\nu\}|H_n, \mathcal{I}) \prod_{\nu=1}^{n} N_\nu^{-\frac{L}{2}} \\ &\times \quad (2\pi)^{\frac{n}{2}}\, |C|^{-\frac{n}{2}}\, \mathrm{etr} \left\{ -\frac{1}{2} C^{-1} \overline{C}^{(n)} \right\} \end{aligned} \tag{6}$$

The definition of the sample mean $\overline{d_\nu}$ and sample covariance $\overline{C}_\nu$ for group ν as well as that for the global sample covariance $\overline{C}^{(n)}$ are

$$\overline{d_\nu} = \frac{1}{N_\nu}\sum_{i=1}^{N_\nu} d_\nu^i \tag{7}$$

$$\overline{C}_\nu = \sum_{i=1}^{N_\nu}(d_\nu^i - \overline{d_\nu})(d_\nu^i - \overline{d_\nu})^T \tag{8}$$

$$\overline{C}^{(n)} = \sum_{\nu=1}^{n}\overline{C}_\nu \tag{9}$$

To avoid clumsy expressions the normalization of the covariance matrices has again been chosen different from the standard form found in text-books. The sample covariances depend on the particular specifications in H_n. The number of indices is minimized as much as possible still allowing an unambiguous interpretation. We see how Occam's factor enters automatically, favoring a small number of groups. The data-driven factors vote for complexity. If the common covariance is known this is the final result and we can start analyzing datasets. Unfortunately, only in very rare cases the covariance matrix is known experimentally and it has to be treated as a hyperparameter.

2.3. POSTERIOR PROBABILITY IF C IS NOT KNOWN

The marginal likelihood $p(D|H_n,\mathcal{I})$ entering Eq.1 is obtained upon marginalization over the true covariance matrix C:

$$p(D|H_n,\mathcal{I}) = \int \mathcal{D}C\, p(D|H_n,C,\mathcal{I})p(C|\mathcal{I})$$

The integral extends over the space of covariance matrices. Fortunately, we can borrow most of the results from what is known about the Wishart distribution [8]. In order to proceed we need the prior $p(C|I)$. Due to the lack of prior knowledge we seek a functional form for the prior which is invariant under all transformations that do not change our state of ignorance. These are all possible similarity transformations. Jaynes' transformation group approach [7] leads to a kind of multivariate Jeffreys' prior

$$p(C|\mathcal{I}) \propto |C|^{-(L+1)/2} \tag{10}$$

The proportionality constant drops out in the odds ratio since it is the same for all hypotheses under consideration. This is the place where the extension to the case of different covariances for different groups needs a little extra thought. The prior (Eq.10) reduces to the well-known Jeffreys' prior in the univariate case. The type of integral, needed for the marginalization in Eq.10, is known from the inverse Wishart distribution [8] and yields

$$P(H_n|D,\mathcal{I}) \propto P(H_n|\mathcal{I})\, P(\{\overline{d}_\nu\}|H_n,\mathcal{I}) \prod_{\nu=1}^{n} N_\nu^{-\frac{L}{2}} \times \pi^{\frac{Ln}{2}} \prod_{j=1}^{L} \Gamma(\frac{N-n-j+1}{2})\, |\overline{C}^{(n)}|^{-\frac{N-n}{2}} \tag{11}$$

This is the key formula to answer the questions we raised in the introduction. It allows to assess preselected groups, and serves as merit function for unsupervised learning (e.g. via simulated annealing). Moreover it allows to address questions like: do we have to split a group? or are there outliers?

2.4. DO WE HAVE TO SPLIT A GROUP?

Here we will consider the question whether the data favor the splitting of a group, η say, into two pieces as depicted schematically in fig. 2. This is summarized in hypothesis H_{n+1}, specifying precisely which data of group η go into the new sub-groups η_1 and η_2. Accordingly, the number of elements change $N_\eta \rightarrow \{N_{\eta,1}, N_{\eta,2}\}$. Eventually we will compare hypotheses H_{n+1} and H_n, with and without the split-

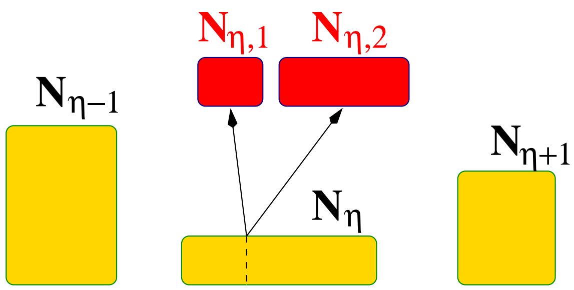

Figure 2. Schematic illustration of the splitting of group η.

ting, which according to Eq.11 reads

$$\frac{P(H_{n+1}|D,\mathcal{I})}{P(H_n|D,\mathcal{I})} = \frac{P(H_{n+1}|\mathcal{I})}{P(H_n|\mathcal{I})} \frac{p(\{\overline{d}_\eta\}|, H_{n+1},\mathcal{I})}{p(\{\overline{d}_\eta\}|H_n,\mathcal{I})} \times \left(2\pi\frac{N_{\eta,1}+N_{\eta,2}}{N_{\eta,1}N_{\eta,2}}\right)^{L/2} |\tilde{C}^{(n)}|^{1/2} \frac{\Gamma(\frac{N-n-L}{2})}{\Gamma(\frac{N-n}{2})} (\frac{N-1}{2})^{L/2} \times \left(1 - \frac{N_{\eta,1}N_{\eta,2}}{(N-1)N_\eta} \Delta d_\eta{}^T (\tilde{C}^{(n)})^{-1} \Delta d_\eta\right)^{-\frac{N-n-1}{2}} \tag{12}$$

Here we switch to the standard definition of the covariance $\tilde{C}^{(n)} = \overline{C}^{(n)}/(N-1)$ and write the difference in sample means of the new sub-groups as $\Delta d_\eta = \overline{d}_{\eta,2} - \overline{d}_{\eta,1}$.

The odds ratio contains the multivariate Student-t distribution (Hotelling-T) which depends essentially on the Mahalonobis distance $\Delta d_\eta{}^T (\tilde{C}^{(n)})^{-1} \Delta d_\eta$ [9].

3. Applications

In order to get a feel for the behavior of Eq.12 we apply it to simple test examples.

3.1. UNIVARIATE EXAMPLE

We start out with two sets of data with 100 data points each, drawn from unit-variance normal distributions with means m_1 and m_2, respectively. Now we want to check whether the bayesian approach recognizes that the data actually belong to two different groups. To this end we calculate the odds ratio for the two hypotheses there are two groups (H_2) versus there is only one group (H_1). To make the test more realistic, we pretend not to know the common noise level and to have equal prior expectation for the two hypotheses,i.e. $P(H_n) = 1/2$. For simplicity we use in the present example a flat prior for $\{m_\nu\}$, i.e. $p(\{m_\nu\}|\mathcal{I}) = \prod_\nu \{\frac{1}{V}\theta(|m_\nu| - \frac{V}{2})\}$. V represents the prior uncertainty in the means m_ν. The data are generated for $m_1 = 0$ fixed and m_2 varied between -0.2 and 0.2. Actually, the result depends only on the difference $m_2 - m_1$. For the numerical analysis we use an prior uncertainty $V = 1$. Fig. 3 shows the various parts of the log-odds-ratio

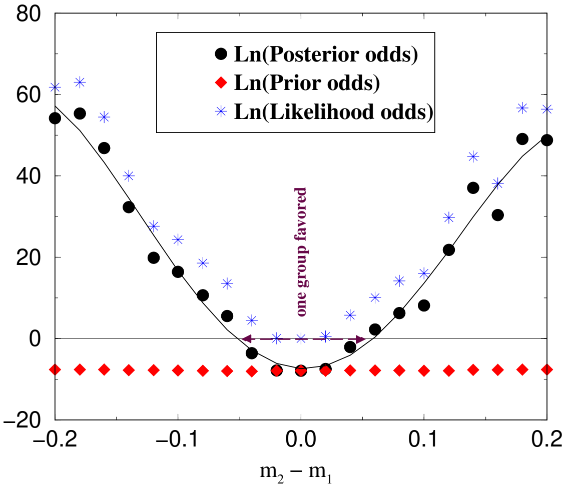

Figure 3. Contributions to the odds ratio for a mock dataset as described in the text, as function of the difference in the true means. Note that posterior and likelihood differ only by a constant prior.

$\log\{P(H_2|D,\mathcal{I})/P(H_1|D,\mathcal{I})\}$, namely prior, likelihood and posterior. In the region of positive log-odds the bayesian approach favors decisively – by several orders of magnitude – two groups if the difference in the true means exceeds the break-even point $\Delta m_{be} = 0.055$. The analytic expression for the break-even point according to Eq.12 yields for $N \gg 1$:

$$(\Delta m_{be})^2 = \frac{4\sigma^2}{N}\log(\frac{NV^2}{8\pi\sigma^2}) \tag{13}$$

As a matter of fact, the data belong to only one group if $m_2 - m_1 = 0$. However, due to the finite sample size and the statistical error in the data, the sample means have an uncertainty which is reflected in the finite value of the break-even point. If on the other hand N – the size of the two samples – tends to infinity, the sample means approach the true means and the break-even point shrinks towards zero.

3.2. OUTLIER DETECTION

An interesting application of Eq.12 concerns outlier detection. This is a special case of Eq.12, namely group η is split into: $N_{\eta,1} = 1$ and $N_{\eta,2} = N_\eta - 1$. A numerical example is studied which consists of 3 groups with 100 data points each. The data of the three groups are distributed according to unit-variance normal distributions with means $m_1 = 2$, $m_2 = 3$ and $m_3 = 4$, respectively. The position of an extra

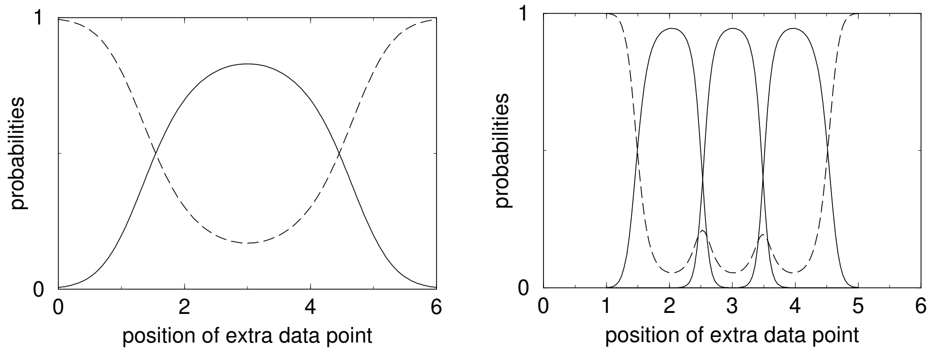

Figure 4. Outlier analysis: data belong to one group (left) or three groups (right) as solid lines. The outlier probabilities are depicted as dashed lines.

test-data point is varied between 0 and 6 and the probabilities are depicted in fig. 4 as function of the position of the test-data point. First, we consider the hypothesis "*all data belong to one group*" and ask for the probabilities that the extra point belongs to that group or that it is an outlier. The result is depicted in the left panel. It corroborates qualitatively our intuition, the outlier probability increases monotonically away from the center of the group. On top of that, Bayesian analysis indicates a failure of the assumed grouping, since the outlier probability is always rather high, even in the center of the group. Alternatively, the probability that the test-data point belongs to the group is always significantly smaller than 1, even for the "best member" of the group. The situation changes if we start from hypothesis H_3 "*there are three groups*". In this case the minimum outlier probability is much smaller. The right panel of fig. 4 shows the probabilities that the extra point belongs to group 1,2,3 or it is an outlier.

An interesting quantity in this respect is the average outlier probability in group ν

$$\overline{p_\nu} = \frac{1}{N_\nu} \sum_{i \in \text{group}_\nu} p(d_i \text{ is an outlier} | D\mathcal{I}) \tag{14}$$

The results are listed in table 1. We see that the average outlier probability is large in the case of the wrong grouping while it is about 5% for the correct grouping. We

	All In One	Three Separate Groups
Group 1	0.30	0.05
Group 2	0.17	0.06
Group 3	0.30	0.05

TABLE 1. Outlier example

see that the Bayesian analysis reveals many interesting features of the problem.

4. Key difference between the groups

Once the optimal grouping has been found the next interesting question arises as to what discriminates the groups. To this end we consider a linear combination $q = \mathbf{a}^T \mathbf{d}$ of the data points to summarize the key features of a group. $\mathbf{a}$ is the discriminating vector. A linear combination is all one can expect in a linear problem. The idea is illustrated in fig. 5. Two data clouds are represented by ellipses. In the original representation in characteristics x_1 and x_2 they can not be distinguished. Projection on the discriminating vector a_2 reveals the existence of two separate groups. We are seeking that vector $\mathbf{a}$ that discriminates the groups best, i.e. for which the odds ratio $P(H_n|\mathbf{a}, D, \mathcal{I})/P(H_1|\mathbf{a}, D, \mathcal{I})$ reaches a maximum. The conditioning on $\mathbf{a}$ means we consider projected, univariate data $\tilde{d}^i_\nu = \mathbf{a}^T \mathbf{d}^{\mathbf{i}}_\nu$. The maximum odds ratio is obtained for

$$\frac{\overline{C}^{(1)}}{\mathbf{a}^T \overline{C}^{(1)} \mathbf{a}} \mathbf{a} = \left(\frac{N-n}{N-1} \frac{\overline{C}^{(n)}}{(\mathbf{a}^T \overline{C}^{(n)} \mathbf{a})} + \frac{n-1}{N-1} \, 1\!\!1 \right) \mathbf{a} \tag{15}$$

For $N \gg n$ we recover the conventional result[10]:

$$\begin{aligned} B\mathbf{a} &= \lambda \, \overline{C}^{(n)} \mathbf{a} \\ B &= \sum_{\nu=1}^{n} N_\nu (\overline{d}_\nu - \overline{d}))(\overline{d}_\nu - \overline{d}))^T \quad . \end{aligned} \tag{16}$$

The covariance of the sample means (covariance between groups) is denoted by B, while $C^{(n)}$ represents an average covariance within the groups. The eigenvalue λ

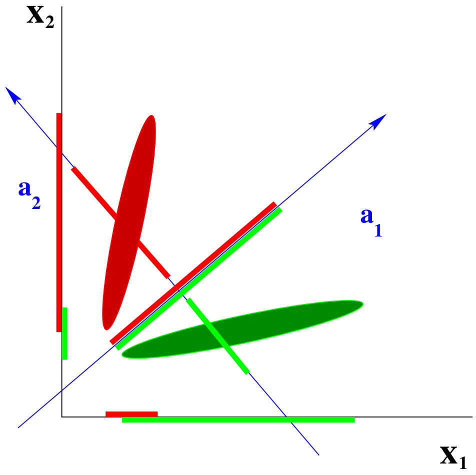

Figure 5. Illustration of the discriminating vector. Ellipses represent clouds of data. Bars along $x_1, x_2, a_1, and a_2$ stand for the projection of the data onto the respective directions.

in Eq.16 is a measure of the ratio $(\frac{\text{distance between the groups}}{\text{distance within the groups}})^2$ for the data projected onto direction **a**. For a sensible grouping the ratio will strongly exceed 1.

5. Fisher's Iris Data

Now we want to apply the theory to a realistic multi-variate problem, namely Fisher's iris data listed and analyzed in [11]. The data consist of 3 pre-formed groups according to the species: iris setosa, iris versicolor and iris virginica. The flowers are characterized only by four measurements: sepal length, sepal width, petal length and petal width. The results are given in table 2. The log-base-10

Choice	$\text{Log}_{10}(P\text{grouping})$	$\overline{P}$(outlier)
(123)	-199	0.41%
(12)(3)	-82	0.15%
(13)(2)	-36	0.12%
(1)(23)	-103	0.19%
(1)(2)(3)	0	0.09%

TABLE 2. Analysis of Fisher's iris data

probabilities (relative to the best model) are given for all different permutations of the pre-formed data groups. The meaning of the notation should be apparent. The integers in parentheses indicate which pre-formed groups are combined to one group. The data corroborate impressively the botanists' classification into 3 separate species. The odds-ratio between the best and next-to-best grouping is 10^{36}! Again we see that the average outlier-probability is much smaller for the best grouping. The smallest value (9%) is, however, about twice as large as in the mock-data set which probably points out that a further sub-classification might be reasonable. The largest eigenvalues of Eq.16 are $\sqrt{\lambda_1} = 5.7$ and $\sqrt{\lambda_2} = 0.53$. There is only one direction along which the groups are well separated. Fig. 6 shows the projected data in the sepal-length – sepal-width subspace (left panel). The data look similar in the other subspaces. The projected data clouds are never separated in the original characteristics. The projection onto the two dominant eigenvectors of Eq.16, however, clearly separates the groups as can be seen in the right panel. The scatter of the data within the groups is much smaller than that between the groups as far as the projection onto a_1 is concerned. The data point enclosed by

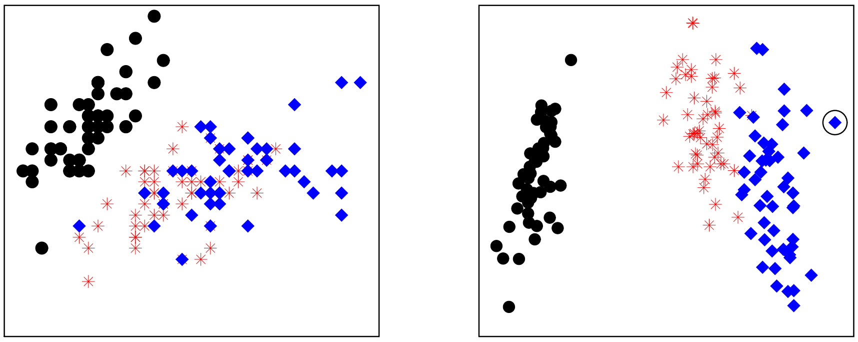

Figure 6. left: sepal width versus sepal length, right:data projected onto the two most dominant directions

a circle is identified to be an outlier.

References

1. W. Krzanowski, ed., *Principles of Multivariate Analysis*, Oxford Science Publications, Oxford, 1988.
2. G. Bretthorst, "On the difference in means," in *Physics and Probability Essays in honor of Edwin T. Jaynes*, W. T. Grandy and P. W. Milonni, eds., vol. 1, pp. 177–194, Cambridge Univ. Press, Cambridge, 1993.
3. J. Stutz and P. Cheeseman, "Autoclass - a bayesian approach to classification," in *Maximum Entropy and Bayesian Methods (1996)*, J. Skilling and S. Sibisi, eds., Kluwer Academic Publishers, Dordrecht, 1995. further references: http://ic-www.arc.nasa.gov/ic/projects/bayesian-group/autoclass/.
4. G. Bretthorst, ed., *Bayesian Spectrum Analysis and Parameter Estimation*, Springer Press, Berlin, Heidelberg, 1988.
5. G. Bretthorst, "Excerpts from bayesian spectrum analysis and parameter estimation," in *Maximum Entropy and Bayesian Methods in Science and Engineering*, G. Erickson and C. Smith, eds., vol. 1, p. 75, Kluwer Academic Publishers, Dordrecht, 1988.
6. G. Bretthorst, "An introduction to parameter estimation using bayesian probability theory," in *Maximum Entropy and Bayesian Methods in Science and Engineering*, P. Fougere, ed., p. 53, Kluwer Academic Publishers, Dordrecht, 1989.

7. E. Jaynes, "Prior probabilities," in *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, R. Rosenkrantz, ed., p. 114, Reidel, Dordrecht, 1983.
8. "Wishart distribution," in *Kendall's advanced theory of statistics, Bayesian Inference*, A. O'Hagan, ed., John Wiley & Sons, New York, 1st ed., 1994. p.293ff.
9. "Mahalonobis distance," in *Principles of Multivariate Analysis*, W. Krzanowski, ed., Oxford Science Publications, Oxford, 1988. p.233ff.
10. "Canonical variates," in *Principles of Multivariate Analysis*, W. Krzanowski, ed., Oxford Science Publications, Oxford, 1988. p.291ff.
11. "Fisher's iris data," in *Principles of Multivariate Analysis*, W. Krzanowski, ed., Oxford Science Publications, Oxford, 1988. p.45ff.