The excess mass approach and the analysis of multi-modality

G. Sawitzki

StatLab Heidelberg, Im Neuenheimer Feld 294, D-69120 Heidelberg

Summary: The excess mass approach is a general approach to statistical analysis. It can be used to formulate a probabilistic model for clustering and can be applied to the analysis of multi-modality. Intuitively, a mode is present where an excess of probability mass is concentrated. This intuitive idea can be formalized directly by means of the excess mass functional. There is no need for intervening steps like initial density estimation. The excess mass measures the local difference of a given distribution to a reference model, usually the uniform distribution. The excess mass defines a functional which can be estimated efficiently from the data and can be used to test for multi-modality.

1. The problem of multi-modality

We want to find the number of modes of a distribution in $\mathbf{R}^k$, based on a sample of n independent observations. There are many approaches to this problem. Any approach has to face an inherent difficulty of the modality-problem: the functional which associates the number of modes to a distribution is only semi-continuous. In any neighbourhood (with respect to the testing topology) of a given distribution, there are distributions with an arbitrarily large number of modes. As a consequence, any confidence interval for the number of modes with finite upper bound will have a confidence level zero (Donoho (1988), Theorem 2.1 and Theorem 2.2).

The impossibility of upper bounds is a combined effect of the semi-continuity, and the richness of the space of probability distributions. If we have restrictions on the family of distributions, upper bounds may be feasible. For example in finite-dimensional parametric families it may still be possible to give non-trivial upper bounds for the number of modes. Unfortunately the restrictions necessary to reduce the space of probability distributions are usually not empirically verifyable. In contrast to problems involving only continuous functionals, with only semi-continuity we cannot even derive approximate solutions for “nearly regular” distributions. Unless we resort to unverifyable assumptions of critical influence, the best we can do is to get **lower bounds** for the number of modes.

Getting lower bounds for the number of modes with guarantied confidence is the first task. Second, we can ask for the power of a procedure. When estimating the number of modes, the challenge is to avoid over-estimation.

GfKI-94 Oldenburg. Printed: Don, 3. Dez 1998

2. The excess mass functional

Any approach has to start with a proper definition of a mode. For a cluster analysis approach, a mode might be defined as a cluster center. For a density estimation based approach, a mode may be identified with a local maximum of the density. In a parametric mixture model, a mode might be related to a mixture component. We try to give here a truly nonparametric approach. Let F be our underlying distribution on $\mathbf{R}^k$. We assume that F has a (bounded, continuous) density f, f>0.

Intuitively, a mode is present where probability mass is concentrated. A large value of the probability density is not enough to guarantee a high mass concentration: a distribution may have isolated spots with high densitiy values, but each with an arbitrarily small support. We may speak of modes of different strengths, depending on the probability mass contained in a mode.

A first step is to measure the mass concentration. Since 'high' mass concentration or 'low' mass concentration are relative properties, we have to take a reference measure. Using a λ-multiple of the Lebesgue measure $\mathbf{R}^k$ as a reference, we define the excess mass at level λ to be the integrated probability mass exceeding the Lebesgue density λ:

$$\boldsymbol{E}(\lambda) = \int (f(x) - \lambda)^+ dx. \qquad (1)$$

with $\mathbf{E}(\lambda) = 1$. At any level λ, the excess mass is the sum of contributions coming from the connectivity components $C_j(\lambda) \subseteq \mathbf{R}^k$ of $\{f \geq \lambda\}$:

$$\boldsymbol{E}(\lambda) = \sum \int_{C_j(\lambda)} (f(x) - \lambda)\, dx. \qquad (2)$$

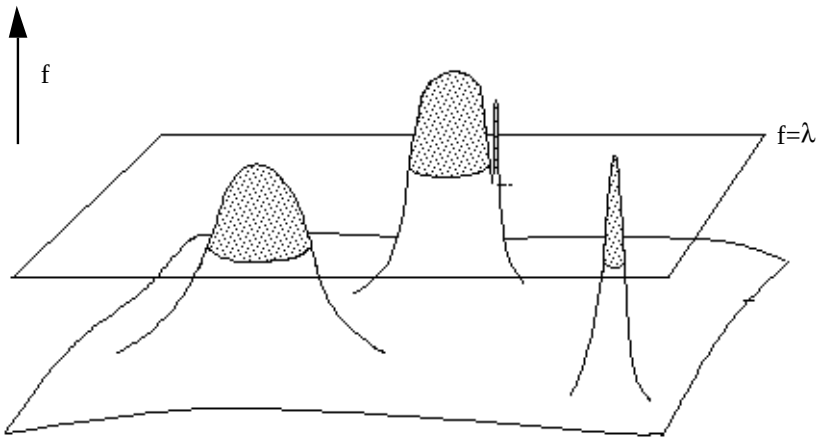

Fig. 1: Density and excess mass. The excess mass is the integrated probability mass exceeding a certain level λ.

For short, the connectivity components $C_j(\lambda)$ of $\{f \geq \lambda\}$ are called λ-clusters. The λ-clusters can be characterized as sets maximizing (2).

$$\boldsymbol{E}(\lambda) = sup_{C \in C_M} \int_C (f(x) - \lambda)\, dx, \tag{3}$$

where $C_M = \{C: C = C_1 \cup C_2 \cup \ldots \cup C_M;\ C_j \subseteq \mathbf{R}^k,\ C_j \neq \emptyset,$ disjoint, connected$\}$ for some $M \geq 1$. This leads to an immediate generalization. For any system of sets C, the excess mass at level λ with λ-clusters in C is defined as

$$E_C(l) = sup_{C \in C} \int_C (f(x) - l)\, dx, \tag{4}$$

with $C = C_M$ as a special case. For a unimodal distribution, at any level λ we have exactly one λ-cluster. For an M-modal distribution, we will have at most M connected components, hence $\mathbf{E}_{C_M}(\lambda) = \mathbf{E}(\lambda)$ for any M-modal distribution F.

Equation (4) has an empirical version. With

$$H_\lambda := F - \lambda \cdot Leb, \tag{5}$$

where Leb is the Lebesue measure in $\mathbf{R}^k$, (4) can be written as $\mathbf{E}_C(\lambda) = \sup_{C \in C} H_\lambda(C)$. Using the empirical distribution function F_n in (5) yields an empirical version

$$H_{n,\lambda} := F_n - \lambda \cdot Leb, \tag{6}$$

leading to an empirical excess mass estimator

$$\boldsymbol{E}_{n,C}(\lambda) = sup_{C \in C}\ H_{n,\lambda}(C). \tag{7}$$

Various assumptions about the modality can be modeled using appropriate choices for C, and tests for multi-modalitiy can be based on the corresponding excess mass estimators. For example, a test for bi-modality can be based on the excess mass difference

$$D_n(\lambda) = E_{n,C_2}(\lambda) - E_{n,C_1}(\lambda), \tag{8}$$

using the maximal excess mass difference

$$sup_\lambda\, D_n(\lambda) = sup_\lambda (\, E_{n,C_2}(\lambda) - E_{n,C_1}(\lambda)\,). \tag{9}$$

as test statistics. Similar tests can be constructed for more general hypotheses and alternatives.

Since for any sets C, C'

$$F_n(C' \backslash C) = 0 \Rightarrow H_{n,\lambda}(C) \geq H_{n,\lambda}(C') \qquad \text{for } C \subset C'$$

and

$$H_{n,\lambda}(C \cup C') = H_{n,\lambda}(C) + H_{n,\lambda}(C') \qquad \text{for } C \cap C' = \emptyset$$

the calculation of the excess mass for usual choices of C amounts to a search for sets in C with components spanned by data points, maximizing (7). In most cases, this is a finite search problem.

3. The excess mass approach

The construction discussed in section 2 is based on the excess mass approach, a general approach which can be applied to a variety of statistical problems (Müller 1992). The basic idea is to find the maximum amount of probability mass which can be attained by a certain model, and to use the exceeding mass as a basis for

further analysis. For the problem of multi-modality, the question is: how much additional probability mass can be attained by a multi-modal model compared to a uni-modal? To answer this question, we have to estimate this excess probability mass $\mathbf{E}_C(\lambda)$ from the data under specific assumptions about the number of modes, e.g.unimodality or bi-modality. To draw our conclusions, we have to study the stochastic behaviour of our excess mass estimator first. Then we can take the estimated excess probability mass as a decision basis. This approach yields diagnostic indices and statistics, which have an immediate empirical interpretation. The decision criterion is the amount of data not fitting a certain model.

While the excess mass approach can be used to find tests or estimators in the classical sense, for many of the practically interesting problems the classical framework is like a procrustean bed. For the multi-modality problem, almost any member of the naïve null hypothesis described by the family of all uni-modal distributions, has most extreme alternatives in any neighbourhood. Defining a useful null hypothesis becomes a problem. The excess mass approach adds to the repertoire as discussed in Gordon (1994). The natural suggestion based on the excess mass approach is to start from the empirical distribution function, find best approximating unimodal models (i.e. distributions minimizing the total variation distance), and to compare the obtained test statistics with the distributions of the excess mass test statistics drawn from these models. As has been pointed out by Davies (1994), this kind of bootstrap fits well into a general framework of data-based inference which explicitly recognizes the approximate nature of probability models.

The excess mass approach has been first applied to the multi-modality problem in Müller and Sawitzki (1987) where the excess mass functional is introduced and first asymptotic results are given for the one-dimensional case. The resulting method is closely related to procedures suggested in Hartigan (1975), Hartigan and Hartigan (1985) and Hartigan (1987).

4. Analysis for multi-modality in one dimension

In one dimension, the situation is simplified, as there is only one choice for the family of possible support sets $\mathcal{C}$. If we have a continuous density, the λ-clusters for an M-modal distribution must be in $\mathcal{C}_M$, the family of sets composed of at most M disjoint intervals. Given a data set, we can explicitely calculate the excess mass for any hypothetical number of modes M by searching for a set composed of at most M intervals with endpoints at data points, maximizing (7).

4.1 Excess mass algorithm in one dimension

The excess mass $\mathbf{E}_{n,M}(\lambda) = \sup_{C\in \mathcal{C}_M} H_{n,\lambda}(C)$ can be calculated stepwise using an iteration over the number of possible modes M. For M=1, this requires the search for an interval with endpoints at data points, i.e. $C_1=\text{argmax } H_{n,\lambda}(C)$. To pass from M to M+1, one of two cases may occur. Additional probability mass may be gained by splitting one of the intervals found in step M (by

removing an open interval with endpoints at data points). Or additional probability mass may be gained by adding an interval in the complement of the intervals found at step M ("splitting lemma" in Müller and Sawitzki (1987)). Both possibilities must be explored, and the maximum contribution taken. The common computational problem resides in finding intervals with maximal ascent (or descent) of $H_{n,\lambda}$. The complexity of this algorithm can be reduced by keeping a "hiker's record list": to find the maximum asscent on your trip, you must keep track of the lowest minimum you have seen so far, and compare the present relative height to the record obtained so far. This gives an algorithm of complexity O(n). More details and an explicit algorithm for the basic search algorithm is given in Müller and Sawitzki (1991).

As a by-product, the algorithm yields the empirical λ–clusters $C_{n,j}(\lambda)$, i.e.. solutions of $\mathbf{E}_{n,M}(\lambda) = \sum_{j=1...M} H_{n,\lambda}(C_{n,j}(\lambda))$, which can be plotted against λ to give a silhouette of the data set. In combination with the excess mass plot, the silhouette can be used for data analysis.

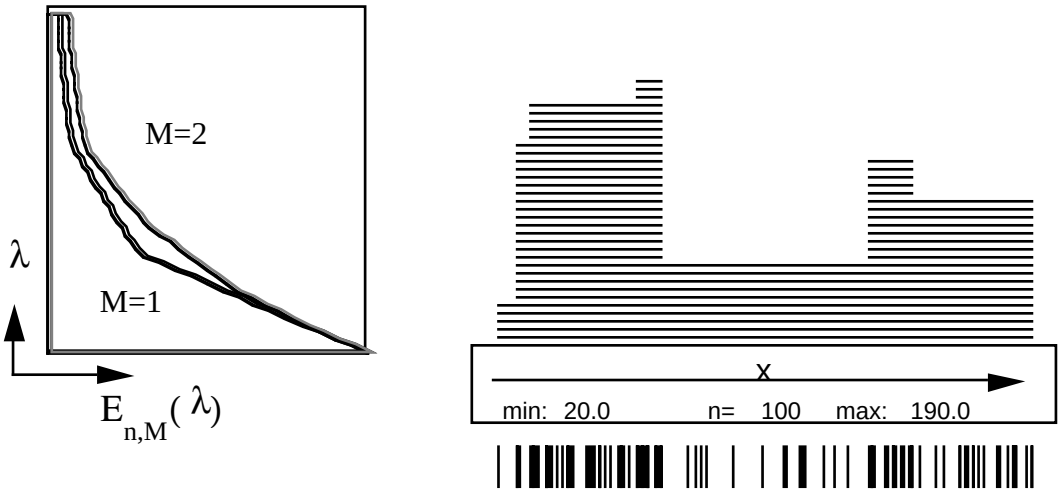

Fig. 2: Estimated excess mass under the assumption of uni-modality (M=1) or bi-modality (M=2) on the left. Silhouette and scatter plot of observed data on the right.

4.2 Asymptotic stochastic behaviour in one dimension

A recommended test-statistics for bi-modality is

$$\boldsymbol{D}_n := sup\ \boldsymbol{D}_n(\lambda) = sup\ \boldsymbol{E}_{n,2}(\lambda) - \boldsymbol{E}_{n,1}(\lambda).$$

More generally we can allow for M modes as an alternative of interest. We have to investigate $\boldsymbol{E}_{n,M}(\lambda)$ under a uni-modal F, but with M>1. Stochastic contributions to the estimation error $\mathbf{E}_{n,M}(\lambda) - \mathbf{E}(\lambda)$ comes from two sources. There is the classical empirical fluctuation governing $H_{n,\lambda} - H_{\lambda}$. And there is an

error since we use estimated λ-clusters instead of the true λ-clusters, hoping that $\{C_{n,j}(\lambda)\} \approx \{C_j(\lambda)\}$. The first asymptotic results can be summarized by:

Theorem (Müller and Sawitzki (1991), Theorem 1):
Let f be a smooth density on **R**, and $x_0 \in \mathbf{R}$ with derivative $f'(x) = 0$ only if $f(x) = 0$ or $x=x_0$. For all $\Lambda > 0$, $M \geq 1$ the process $\lambda \rightarrow \sqrt{n}(\mathbf{E}_{n,M}(\lambda) - \mathbf{E}(\lambda))$ converges weakly in $D[0,\Lambda]$ to $\lambda \rightarrow B(a_\lambda)$, B a standard-Brownian bridge, where $a_\lambda = P_F\{x|f(x)\geq\lambda\}$.

This theorem guarantees a square root asymptotics for the excess mass estimator under the unimodal hypothesis. This is a better rate than usually is achived. The key is that the excess mass functional contains information about mass concentration, but does not try to identify mass location. Separating the question of mass concentration from location allows a better error rate. Confidence bands can be constructed, using this theorem.

The behaviour of the suggested test statistics $\mathbf{D}_n$ is characterized by
Theorem (Müller and Sawitzki (1991), Theorem 2):
Let f be unimodal with $f'(x) = 0$ iff $f(x) = 0$ or $x=x_0$;
f' ultimately monotone in the tails;
f'' bounded in a neighbourhood of x_0, with $f''(x_0) < 0$.
Under these conditions:

(i) $\boldsymbol{D}_n(f(x_0)) = \boldsymbol{O}_P(n^{-3/5})$

(ii) $max_{\lambda \leq f(x_0)-\varepsilon} \boldsymbol{D}_n(\lambda) = \boldsymbol{O}_P(n^{-2/3}log^{2/3} n)$ $(\varepsilon>0)$

(iii) $max_\lambda \boldsymbol{D}_n(\lambda) = \boldsymbol{O}_P(n^{-3/5}log^{3/5} n)$

This theorem tells that in the one-dimensional situation the essential stochastic contribution to the excess mass difference comes from the mode (3/5< 2/3 !). For the uniform distribution, we would have $\max_\lambda \mathbf{D}_n(\lambda) = \boldsymbol{O}_P(n^{-1/2})$. The difference in order is sizeable: for a sample size of n=50, the difference in order $n^{1/10}$ has a numeric value of 1.47.

5. Analysis for multi-modality in higher dimensions

In higher dimensions, additional difficulties occur. First, the family of possible λ-clusters is an open choice. While in one dimension any disjoint union of intervals are the obvious candidates, we have more freedom of choice in higher dimensions. Second, the tools at hand are restricted. In one dimension, the Komlós-Major-Tusnády machinery could be used to derive the asymptotic behaviour of the empirical excess mass differences. However this does not have an immediate extension to higher dimensions. Instead, empirical process theory must be used which requires a stricter control of the families of sets under discussion.

The choices of basic set families C_M in higher dimensions must be governed by two rationales. They must be sufficiently rich to allow at least for classical mixture models, like the mixture of normal distributions. On the other hand, they must be sufficiently sparse to allow empirical process theory, or allow for an

adequate ad-hoc theory. Usual choices are sparse classes, like Vapnic-Cervonenkis classes, guaranteeing a small coverage dimension, or richer classes, like *conv*^{2}, the convex sets in the plane, as considered in Hartigan (1987).

For any choice of set systems $\mathcal{C}_1 \subset \mathcal{C}_2$ we can define empirical excess mass estimators $\mathbf{E}_{n,\mathcal{C}_1}(\lambda)$, $\mathbf{E}_{n,\mathcal{C}_2}(\lambda)$ as above and use the excess mass difference $\mathbf{D}_n(\lambda) = \mathbf{E}_{n,\mathcal{C}_2}(\lambda) - \mathbf{E}_{n,\mathcal{C}_1}(\lambda)$ to define a test for the hypothesis $\{f \geq \lambda\} \in \mathcal{C}_1$. To test against bimodaltity, $\mathcal{C}_1$ will be chosen to have one connectivity component, and $\mathcal{C}_2$ having two. But other choices modelling qualitative assumptions on the shape of the λ-clusters by appropriate choice of $\mathcal{C}_1$ and $\mathcal{C}_2$ are covered by the same framework (Polonik 1993a).

5.1 Asymptotic stochastic behaviour in higher dimensions

As in the one-dimensional case, a major step is to get hold of the estimation error involved in using an empirical λ-cluster $C_n(\lambda)$ instead of the true set $C(\lambda)$. A key tool is the inequality due to Polonik (1993):

$$Leb\{C(\lambda) \Delta C_n(\lambda)\} \leq Leb\{x: |f(x)-\lambda| < \varepsilon\} + \varepsilon^{-1}\{(F_n-F)(C_n(\lambda)) - (F_n-F)(C(\lambda))\} \quad \forall \varepsilon > 0.$$

This inequality separates analytical properties of the density f (first term) from oscillation of the process F_n-F (second term).

The asymptotic behaviour of the excess mass difference is characterized by the following theorem (Polonik 1993):

Theorem Let f be regular unimodal density (i.e. elliptical at mode x_0 + regularity+ rapidly decreasing tails). Then

(i) if $\mathcal{C}_2$ is a VC-Class:

(dimension 1) $\quad max_\lambda \mathbf{D}_n(\lambda) = \mathbf{O}_P(n^{-3/5} log^{3/5} n)$

(dimension > 1) $\quad max_\lambda \mathbf{D}_n(\lambda) = \mathbf{O}_P(n^{-2/3} log^{2/3} n)$

(ii) if $\mathcal{C}_2$ consists of finite unions of differences in conv^2:

$$max_\lambda \mathbf{D}_n(\lambda) = \mathbf{O}_P(n^{-4/7}).$$

In contrast to the one-dimensional situation, for any dimension > 1 there is no general dominating contribution from the modes since $Leb\{x:|f(x)-f(x_0)| < \varepsilon\} \approx \varepsilon^{1/2}$ for dimension one, but $Leb\{x:|f(x)-f(x_0)| < \varepsilon\} \approx \varepsilon^p$ with $p \geq 1$ in higher dimensions.

The excess mass difference for a uniform distribution on a bounded region has rates $\mathbf{O}_P(n^{-1/2})$, hence for VC-classes: the previous exponents differ at most by 1/6 (for illustration: $50^{1/6}$=1.919...).

5.2 Excess mass algorithms in higher algorithms

While the general algorithmic approach sketched above still holds in higher dimensions, general effective algorithms are not available in higher dimensions. The search space is defined by the choice of the model spaces $\mathcal{C}_M$. For convex sets in two dimensions, the algorithm suggested by Hartigan (1987) can be

applied. For ellipsoids, Nolan (1991) uses a variant of the Rousseuw and Leroy algorithm for minimal volume ellipsoids. Nason and Sibson (1992) suggest a combination of lower dimension search strategies with approaches from projection pursuit, like the grand tour method. But so far too little is known about appropriate search algorithms which can be applied here.

6. Tests for multi-modality

Despite the detailed asymptotics, the finite sample distribution of the excess mass difference is not yet sufficiently known. We can see three approaches to derive valid tests.
First, we can derive stochastic bounds. In one dimension, these bounds can be based on

$$sup_\lambda D_n(\lambda) \le max_C |(F_n - F)(C)| \qquad (10)$$

The right hand side is well-understood in one dimension (Müller and Sawitzki, 1991). Unfortunately this bound appears to be very conservative. A similar bound is possible in higher dimensions (Polonik 1993a).

Second, we can derive critical values from special model distributions. For one dimension, sample size n=50 and a Gaussian, Cauchy and uniform model distribution, the resulting distribution of the test statistics is plotted in (Müller and Sawitzki, 1991). For the uniform distribution, as an extremal case of unimodal distributions, the distribution is tabulated in (Müller and Sawitzki, 1991).

Third, we can bootstrap the excess mass difference based on the estimator $f_n(x) = \max\{\lambda \ge 0 : x \in C_{n,1}(\lambda)\}$ as an estimator of the best-approximating unimodal distribution. Consistency and quality of this bootstrap approximation however still need further investigation.

References:

DAVIES, L. (1994): Data features. Manuscript. Essen 1994. To appear in *Statistica Nederlandica.*

DONOHO, D.L. (1988): One-sided inference about functionals of a density. *The Annals of Statistics, 16, 1390-1420.*

GORDON, A.D. (1994): Null models in cluster valiadation. In: W. Gaul, D. Pfeifer (eds.) From data to knowledge: Theoretical and practical aspects of classification, data analysis and knowledge organization. Proc. 18th Annual Conference of the GfKl, Univ. of Oldenburg, 1994. Springer Verlag, Heidelberg Berlin, 1994 (in preparation)..

HARTIGAN, J.A. (1975): Clustering algorithms. Wiley, NewYork.

HARTIGAN, J.A., and HARTIGAN, P.M. (1985): The dip test of unimodality. *Annals of Statistics, 13, 70-84.*

HARTIGAN, J.A. (1987): Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association, 82, 267-270.*

MÜLLER, D.W., and SAWITZKI, G. (1987): Using excess mass estimates to investigate the modality of a distribution. Preprint Nr. 398, Januar 1987, Universität Heidelberg, Sonderforschungsbereich 123 Stochastische Mathematische Modelle.

MÜLLER, D.W., and SAWITZKI, G. (1991): Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association, 86, 738–746.*

MÜLLER, D.W. (1992): The excess mass approach in statistics. *Beiträge zur Statistik 3. ftp: statlab.uni-heidelberg.de*

NASON, G.P., and SIBSON,R. (1992): Measuring multimodality. *Statistics and Computing 2, 153-160.*

NOLAN, D. (1991): The excess-mass ellipsoid. *Journal of Multivariate Analysis, 39, 348-371.*

POLONIK, W. (1993): Measuring mass concentration and estimating density contour clusters – an excess mass approach. *Beiträge zur Statistik 7. ftp: statlab.uni-heidelberg.de.* Submitted to *Annals of Statistics.*

POLONIK, W. (1993a): Density estimation under qualitative assumptions in higher dimensions. *Beiträge zur Statistik 15. ftp: statlab.uni-heidelberg.de*