The Excess Mass Approach in Statistics

D. W. Müller

Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg, Germany
e-mail: dwm@statlab.uni-heidelberg.de

3

Universität Heidelberg
Institut für Angewandte Mathematik
Im Neuenheimer Feld 294
69120 Heidelberg
Germany

Beiträge zur Statistik
Nr. 3
December 1992

The Excess Mass Approach in Statistics

D.W. Müller
Institut für Angewandte Mathematik,
Universität Heidelberg

The basic idea of the excess mass approach is to measure the amount of probability mass not fitting a given statistical model. It came up first in the context of testing for a treatment effect, was later applied to inference about the modality of a distribution and even density estimation. Recently the framework has been extended to regression problems.

The motivation behind this line of research is an old one. All statistical methods rely on assumptions about the underlying statistical models. Some of these assumptions cannot be verified on the basis of empirical observations. Research has been done to weaken those assumptions for known methods or even to develop new statistical procedures that work under minimal assumptions. Thus, during the past decades, resampling procedures have been developed that turn out to work under less severe assumptions than classical methods.

(Some remarkable work in this direction has been done in this Sonderforschungsbereich by Enno Mammen who could establish the superiority of certain resampling methods when the parameter dimension is high (e.g. Mammen (1989, 1992)). A different approach consists in developing diagnostic procedures that detect violations of the assumptions. Much of the work of Werner Ehm done in this Sonderforschunsbereich has been devoted to this problem, mainly in the context of frequency data (Ehm (1991)).

The excess mass approach is a different way of dealing with the problem of hidden nuisance parameters. It restricts attention to a special class of statistics that do not depend too heavily on the underlying parameters. The nature of this approach will be described here.

---

[1] Presented as the closing lecture in the final colloquium ("Abschlusskolloquium") of the Sonderforschungsbereich 123 "Stochastische Mathematische Modelle" in Heidelberg, December 12, 1992.

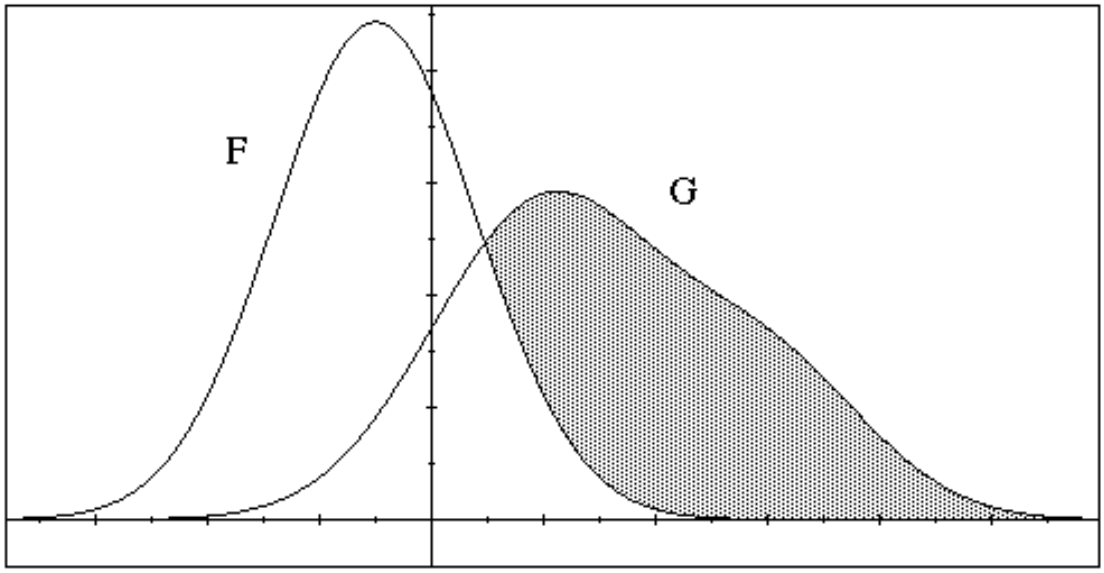
1. The origin - a testing problem

Let F, G be distributions of real-valued observables X, Y respectively describing the situation before and after treatment, with inference about the joint distribution of X and Y being impossible. (This is often the case when the object under observation is altered, or even destroyed, in the experiment). Inference will be based on independent samples $x_1,...,x_n$ from F and $y_1,...,y_n$ from G. Has there been a treatment effect?

The traditional formulation of this problem is strongly model-dependent. It formulates a hypothesis that the combined sample comes from the same distribution F = G, and then derives tests under model assumptions about F and G. Considerable effort has been taken to construct tests that are robust against violations of those assumptions. Still the success is not satisfactory.

An entirely different approach consists in *measuring* the treatment effect (Müller (1980)). This can be done in the following way. We model the treatment process as a Markov transition kernel K_x, the distribution of Y given X = x, such that $K_x[Y \geq x] = 1$ for all x. This kernel refers to a latent structure, the joint distribution of X,Y, which is not identifiable in our setup. The condition imposed on K just means that there cannot be negative effects. A measure of the size of the treatment effect would be the proportion of the population that benefitted from the treatment. This is a quantity that depends on the joint distribution of X and Y, and thus is not identifiable.

We therefore consider the *minimal guaranteed* treatment effect, given F and G, i.e. the minimal proportion π of the population that benefitted from the treatment. Clearly
$\pi = \pi(F,G) = \min_K Ws[Y > X]$. This is a marginal quantity, depending on F and G alone. It can be shown that $\pi(F,G) = \|(G-F)^+\|_1$. The picture shows this quantity as the content of the shaded area.

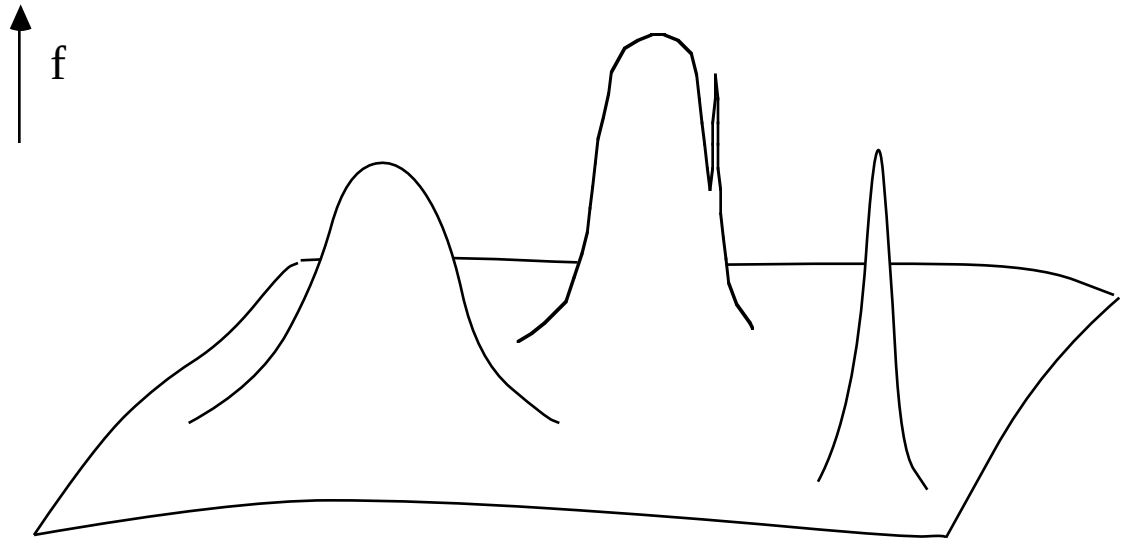


It would be a natural procedure to estimate $\pi(F,G)$ by plugging in the empirical distributions F_n,G_n of F, G respectively. This does not work since, in general, F_n and G_n will be disjoint measures. The way out is to observe that the L^1-norm is a sup over the class of *all* measurable sets C of the real axis such that $\pi(F,G) = \sup_C(F\text{-}G)(C)$. This class being too large we model a smaller class $\mathcal{C}$ of possible supports of $(G\text{-}F)^+$, assuming that $\pi(F,G) = \sup_{C\in\mathcal{C}} (F - G)(C)$. In this form we obtain an estimator of $\pi(F,G)$ by a plug-in: $\pi(F_n,G_n) = \sup_{C\in\mathcal{C}} (F_n - G_n)(C)$. E.g. one can take $\mathcal{C}$ as the class of all intervals of the form $[t, +\infty)$ (t real), provided that dG/dF is monotonically increasing. In this case $\pi(F_n,G_n) = \sup_t(G_n(t) - F_n(t))$. Thus $\pi(F_n,G_n)$ is the well-known Kolmogorov-Smirnov statistic. More general cases have been treated in Müller (1980).

The quantity $\|(G\text{-}F)^+\|_1$ will be called the "excess mass of G over F". Here F is considered as a "reference measure".

2. New looks at old problems - investigating multimodality

The new way of thinking - namely using reference measures for statistical inference - has led to new formulations also for other statistical problems. The problem which we have looked at first concerns the inference about the *number of modes* of a distribution.



Classical methods in this field work via an initial density estimation. They assume a smooth density in the background and think of modes as zeros of its derivative. In contrast to this, we *measure* the modality by *measuring mass concentration*. Thus we base the statistical analysis on functionals which reflect the "distinctiveness" of a mode. The literature already contains a number of related proposals. We just recall a few concepts which have been introduced in the literature: “modal intervals” by Lientz (1970), “modes with given width” by Hartigan (1977), “bumps” by Good & Gaskins (1980).

Our approach can be briefly described as follows: let F be a probability distribution on $\Re^d$, with a continuous density f. The reference measures to which f will be compared are multiples of Lebesgue measure. Thus the “excess mass” will be

$$\begin{aligned} \mathbf{E}(\lambda) &= \|(F - \lambda \cdot \mathrm{Leb})^+\|_1 \\ &= \int (f(x) - \lambda)^+ dx \end{aligned}$$

where λ is a free parameter controlling the size of mass concentration. Historically the first references are Müller (1981), Müller & Sawitzki (1987), Müller & Sawitzki ICOSCO **Ç**esme (1987), Hartigan (1987).

It will be noted that

$$\mathbf{E}(\lambda) = \int_{\cup I_j(\lambda)} (f(x) - \lambda)\, dx,$$

where $I_j(\lambda)$ are the connected components of $[f \geq \lambda]$ ("λ-clusters") (cp. “density contour

clusters" introduced by Hartigan(1975)). Thus, for f with at most M modes $\mathbf{E}(\lambda)$ can be written as

$$\sup \int_{\cup I_j(l):\ j=1,\ldots,M} (f(x) - \lambda)\, dx,$$

where the sup is extended over I_j, j = 1,…,M, pairwise disjoint connected sets, i.e.

$$\mathbf{E}(\lambda) = \sum_{j=1,\ldots,M} (F - \lambda\, \mathrm{Leb})(I_j),$$

a form suitable for estimation by the plug-in method!

3. Concept

The general framework can be described thus:

there is a class $\mathcal{C}$ of (measurable) subsets of $\Re^d$ (the "support" class),
and it is assumed that for the underlying density f one has $C(\lambda) \equiv [f \geq \lambda] \in \mathcal{C}$ $(\lambda \geq 0)$.

(There is a development which works without this assumption, see Polonik (1992); the complications to which it leads do not make it look suitable for presentation here).

The following quantities will be of interest.

(a) The level measure $H_\lambda = F - \lambda \cdot \mathrm{Leb}$
and its empirical counterpart $H_{n,\lambda} = F_n - \lambda \cdot \mathrm{Leb}$,
where F_n denotes the empirical measure.

(b) The excess mass functional $\mathbf{E}(\lambda) = \sup_{C \in \mathcal{C}} H_\lambda(C)$
and its empirical counterpart $\mathbf{E}_n(\lambda) = \sup_{C \in \mathcal{C}} H_{n,\lambda}(C)$.

Then $\mathbf{E}(\lambda) = H_\lambda(C(\lambda))$ and $\mathbf{E}_n(\lambda) = H_{n,\lambda}(C_n(\lambda))$ where $C_n(\lambda) = \mathrm{argmax}\, H_{n,\lambda}$.

3.a Example (d=1)

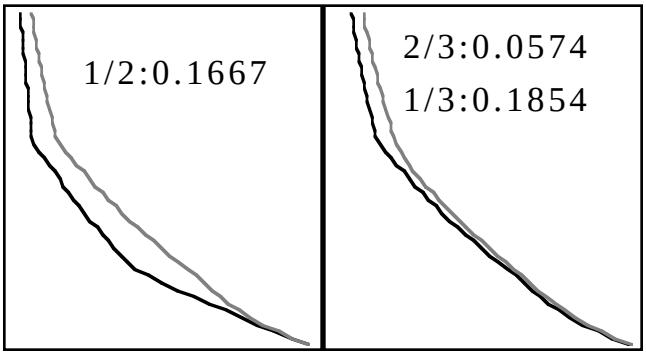
Let $C = C_M$ be the family of all unions of at most M intervals of the line.

For testing unimodality versus bimodality one considers the excess mass difference $\mathbf{D}(\lambda) = \mathbf{E}_{C_2}(\lambda) - \mathbf{E}_{C_1}(\lambda)$, in particular $\max_\lambda \mathbf{D}(\lambda)$. For a bimodal distribution F this is just the total variation difference to the closest unimodal distribution. With its empirical counterpart $\mathbf{D}_n(\lambda) = \mathbf{E}_{n,C_2}(\lambda) - \mathbf{E}_{n,C_1}(\lambda)$ this suggests $\max_\lambda \mathbf{D}_n(\lambda)$ as test statistic for unimodality versus bimodality.

Here is a data example.

Eruption length (in minutes) of 107 eruptions of Old Faithful Geyser
(Source: Silverman (1986), Table 2.2)

For these data, the curves $\mathbf{E}_{n,C_1}(\lambda)$ versus $\mathbf{E}_{n,C_2}(\lambda)$ and $\mathbf{E}_{n,C_2}(\lambda)$ versus $\mathbf{E}_{n,C_3}(\lambda)$ are plotted in the following figure. The value assumed by $\max_\lambda \mathbf{D}(\lambda)$ is 0.1667.



Excess mass curves $\mathbf{E}_{n,C_1}(\lambda)$, $\mathbf{E}_{n,C_2}(\lambda)$ and $\mathbf{E}_{n,C_3}(\lambda)$, for the Old Faithful Geyser Data

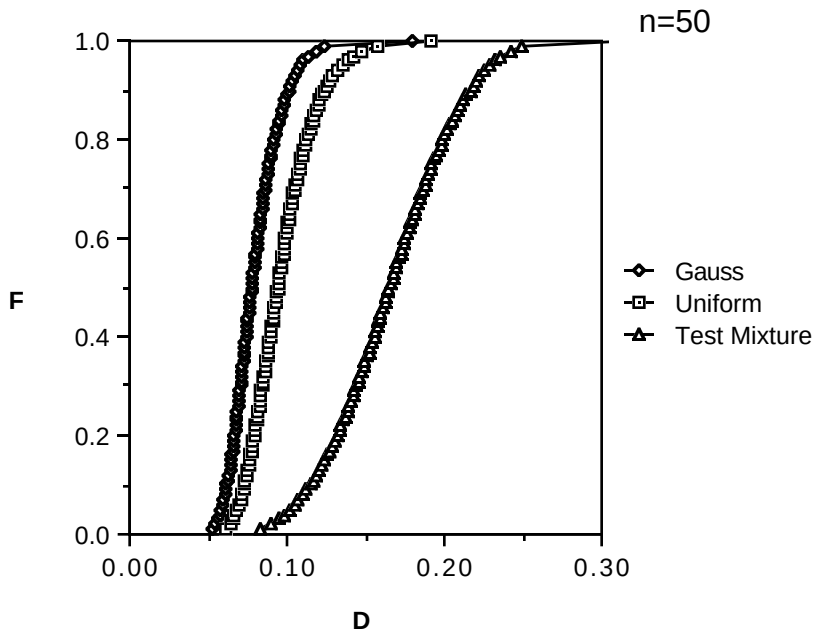
It is desirable to infer the true value of $\max_\lambda \mathbf{D}(\lambda)$ from the observed value of $\max_\lambda \mathbf{D}_n(\lambda)$. This is an unsolved problem! Since nothing is known about the distributional properties of this quantity so far, attention should be restricted to simpler questions first, such as this: is

the observed value a large one, possibly giving hints that the underlying distribution is indeed multimodal?

For judging the observed value, the "zero-effect" of the statistic should be known. This means the distribution of the statistic should be studied under some unimodal distributions, and the observed value should be compared to these distributions.

(At first sight, this procedure might resemble what is called hypothesis testing in the decision-theoretic formulation of statistics. As already mentioned above, it is of an entirely different nature and avoids the well-known drawbacks of the decisions-theoretic approach. In statistical practice one rarely fixes the level of a test in advance and does not exclude the possibility of accepting a hypothesis. Rather one wants to measure the degree of violation of a hypothesis (e.g., see Martin-Löf (1974)). In contrast to the classical formulation, our statistic is not derived as the result of some optimality problem, but introduced by the need for giving questions like "how many cases fit the hypothesis" a mathematical formulation. Moreover, really, no decision is intended. Rather, the need for a quantitative support of the statisticians´s judgement is widely felt.

In our situation of the multimodality problem the simplest question is, whether the observed value of the statistic should be judged large in comparison with the values normally occurring in standard unimodal situations. Simulation results for such situations would be a first step. Such results are reported in Müller & Sawitzki (1991) for the Cauchy, Gauss, and uniform distributions. The results for the Gauss and uniform distribution and a bimodal mixture of two uniform distributions are displayed in the exhibit. In all simulation runs the uniform case has proved to be the stochastically largest. As follows from Hartigan & Hartigan (1985) this cannot be true in general, however! It is expected that some asymmetric "almost bimodal" distributions (e.g. certain bi-uniforms) could prove to be counterexamples. While such a behaviour seems to be restricted to "pathological" cases, it is legitimate to take the uniform as a conservative "standard").



Simulated distribution functions under the Gauss and uniform distributions, and under a bimodal test distribution

For the present data the empirical excess mass difference 0.1667 is well above the 99% - quantile, calculated under the uniform distribution, thus strongly supporting the hypothesis of multimodality.

3.b Mathematical explanation

Under bimodal distributions, the excess mass difference statistic tends to be larger than under unimodal distributions. No mathematical statement is known making this sufficiently precise. Still, there is a weak explanation in terms of asymptotic rates. There is a difference in rates between the “regularly” unimodal, the uniform and the multimodal situations. The following theorem refers to the “regularly” unimodal case.

Theorem (Müller & Sawitzki (1991)).
Let the density f *of* F *be unimodal with* $f'(x) = 0$ *only if* $f(x) = 0$ *or* $x=x_0$. *Let* f' *beultimately monotone in the tails and* f''' *bounded in a neighborhood of* x_0, *with* $f''(x_0) < 0$.

Then

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

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

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

As (i) shows in contrast with (ii), the essential contribution to $\max_\lambda \mathbf{D}_n(\lambda)$ comes from the mode ($3/5 < 2/3$). The rate is due to the elliptical behaviour of the density near the mode. It varies with the degree of flatness and becomes $\max_\lambda \mathbf{D}_n(\lambda) = \boldsymbol{O}_P(n^{-1/2})$ in the extreme case of the uniform distribution. The slower rate can partially explain why $\max_\lambda \mathbf{D}_n(\lambda)$ appears stochastically largest among the standard unimodal distributions considered. On the other hand, the difference in order is $n^{1/10}$, small enough for decent sample sizes ($50^{1/10} = 1.47...$) to provide a partial justification for considering the uniform case as a conservative unimodal "standard". Clearly, for multimodal distributions, $\max_\lambda \mathbf{D}_n(\lambda)$ stays away from zero.

It has turned out that $\max_\lambda \mathbf{D}_n(\lambda)$ is equivalent to Hartigan´s DIP-statistic. Therefore the above result can be regarded as an extension of the asymptotic results of Hartigan & Hartigan (1985).

3.c The general case (d ≥ 1)

We consider two classes of sets $C_1 \subset C_2$ such that $[f \ge \lambda] \in C_1$ for all real λ. In this general setup the excess mass difference process is $\lambda \to \mathbf{D}_n(\lambda) = \mathbf{E}_{n,C_2}(\lambda) - \mathbf{E}_{n,C_1}(\lambda)$. While for d = 1 the only interesting support classes are made up of intervals, the choice will be more delicate in higher dimensions. The basic distinction emerges from empirical process theory: there are

(i) "poor" classes like Vapnik-Cervonenkis-classes
(with small covering dimension);

(ii) "rich" classes like $conv^2$, the class of all convex subsets of the plane
(this case has found special attention in Hartigan (1987)).

The results for d = 1 were obtained essentially via Hungarian embeddings, which are not available for larger d. For higher dimension Polonik(1992) has developed a completely different method. He observed that the problem can be reduced to estimating the size of $C_{discrep}(\lambda) \equiv C(\lambda)\Delta C_n(\lambda)$. Here is Polonik´s inequality:

$$Leb\{C_{discrep}(\lambda)\} \leq Leb\{x: |f(x) - \lambda| < \varepsilon\} + \varepsilon^{-1}\{(F_n - F)(C_n(\lambda)) - (F_n - F)(C(\lambda))\}$$

This inequality accentuates *analytical properties of the density* f (1^{st} term) and the *oscillation of the empirical process* F_n-F (2^{nd} term) and separates both factors.

This is Polonik´s **proof.**
First one notes that H_λ has density f - λ. Then one considers the integral

$$\int_{C_{discrep}(\lambda)} |f - \lambda| \, dx = H_\lambda(C(\lambda)) - H_\lambda(C_n(\lambda))$$

$$= H_{n,\lambda}(C(\lambda)) - (F_n - F)(C(\lambda)) - H_{n,\lambda}(C_n(\lambda)) + (F_n - F)(C_n(\lambda))$$

$$\leq 0 + (F_n - F)(C_n(\lambda)) - (F_n - F)(C_n(\lambda)).$$

Obviously, a lower bound of the integral is $\varepsilon\, Leb\{C_{discrep}(\lambda) \cap \{x: |f(x)-\lambda| \geq \varepsilon\}\}$.
Finally one decomposes $C_{discrep}(\lambda)$ according as $|f(x) - \lambda| \geq \varepsilon$ or $|f(x) - \lambda| < \varepsilon$.■

With the help of this inequality, Polonik arrives at the following theorem.

Theorem (Polonik (1992)).
Let f *be regularly unimodal (i.e.elliptical at the mode* x_0 *, rapidly decreasing in the tails, and satisfying further regularity conditions). Then*

(i) ($\mathcal{C}_2$ *a Vapnik-Cervonenkis class*)

$$(d = 1) \qquad \max_\lambda \mathbf{D}_n(\lambda) = \boldsymbol{O}_P(n^{-3/5}\log^{3/5} n),$$
$$(d > 1) \qquad \max_\lambda \mathbf{D}_n(\lambda) = \boldsymbol{O}_P(n^{-2/3}\log^{2/3} n).$$

(ii) ($\mathcal{C}_2$ *the family of finite unions of differences of conv*^{2})

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

It is noteworthy that for d > 1 there is no essential contribution by the modes:

$$\text{Leb}\{x: |f(x) - f(x_0)| < \varepsilon\} \approx \varepsilon^{1/2} \ (d = 1), \approx \varepsilon^p \text{ with } p \geq 1 \ (d > 1).$$

Again, the (multivariate) uniform distributions supported by bounded regions represent separate cases yielding special rates $\boldsymbol{O}_P(n^{-1/2})$ so that for Vapnik-Cervonenkis-classes the exponents differ by at most 1/6 ($50^{1/6}$=1.919...).

4. Unfoldment

The idea of excess mass estimation has given rise to various extensions and new developments such as

(a) Excess mass ellipsoid estimation (Nolan (1992)) - this is the problem of estimating $C(\lambda)$ by $C_n(\lambda)$ in the class $\mathcal{C}$ of all ellipsoids;

(b) Density contour estimation - this is the problem of nonparametric estimation of $C(\lambda)$ in general classes of sets (Polonik (1992));

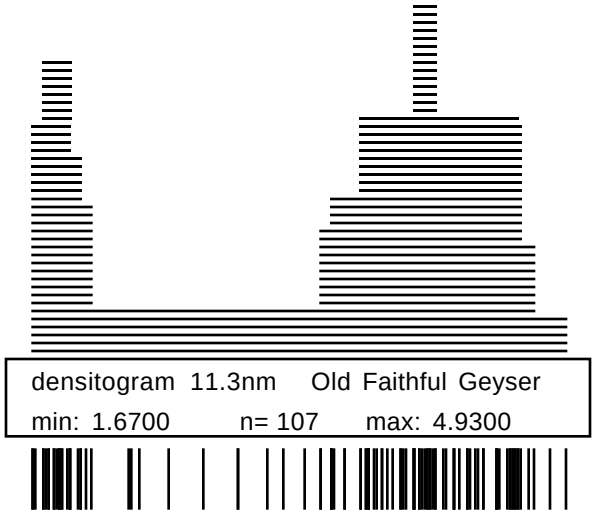
(c) Density estimation - to explain the relation to density estimation we write f(x) as

$$f(x) = \int_{x \in C(\lambda)} d\lambda \ .$$

For a given bound M (assuming $f \le M < \infty$) we form the plug-in-estimator

$$f^*_n(x) = \int_{x \in C_n(\lambda), \lambda \le M} d\lambda$$

(called "silhouette" by Müller & Sawitzki (1987)). For the Old Faithful Geyser data a grafical display of this estimator is shown in the figure.



The silhouette for the Old Faithful Geyser data.

(c.i) Relatives

The silhouette estimator f^*_n turns out to be a natural generalization of known nonparametric devices. It reduces to the Grenander estimator for monotone densities, d=1 (Polonik (1992)).

(c.ii) Rates

Thus the theory developed in Polonik (1992) naturally extends the asymptotic results obtained for the Grenander estimator (e.g. Groeneboom (1985)). Let $\mathcal{C}$ be a class of sets such that $[f \ge \lambda] \in \mathcal{C}$ for all real λ, and consider the function $\Psi(t) = \int \min(f(x),t)\, dx$. Then

(i) $f^*_n(x) - f(x) = \boldsymbol{O}_P(\Psi(n^{-1/3} \log^{1/3} n))$ in L^1

(for certain "regular" Vapnik-Cervonenkis-classes $\mathcal{C}$ - see Polonik (1992)) and

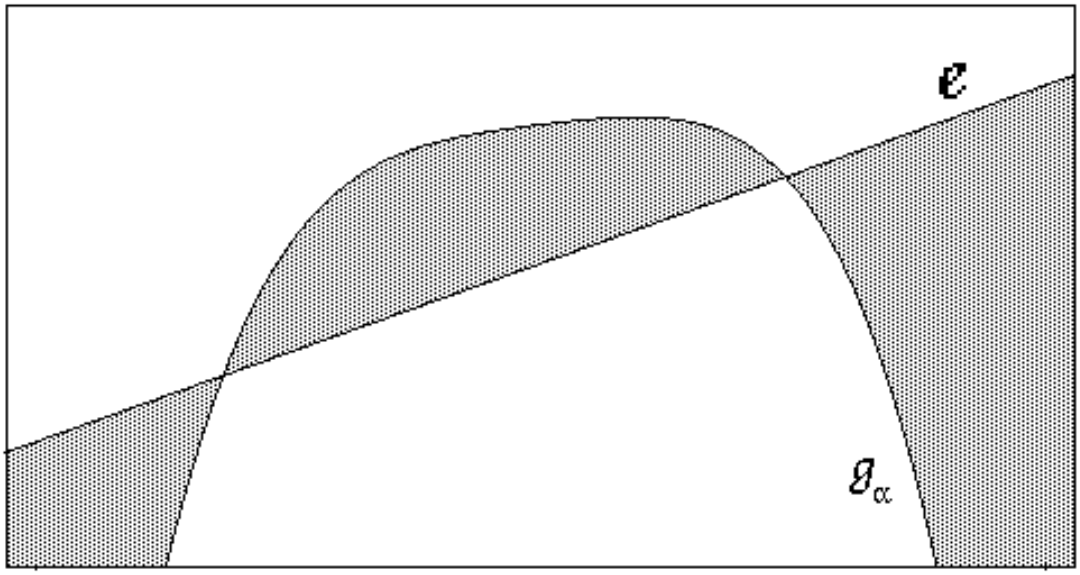
(ii) $f^*_n(x) - f(x) = \boldsymbol{O}_P(\Psi(n^{-2/7}))$ in L^1 (for $\mathcal{C} = conv^2$).

(d) Regression

Is there an excess mass analogue in regression? We approach this question by asking, for a given regression function, how much probability mass will not fit this model. Suppose we are looking for the conditional α-quantile function g_α of a distribution P on the space $\Re^d \times \Re^1$, i.e.

$P[y < g_\alpha | x] = \alpha$ for all $x \in \Re^d$.

To fix ideas, we suppose here d=1 and look for a *linear fit* of g_α (for more general models see (Müller (1992)). The model of a linear regression function not being true exactly, let us assume it to be *concave*, say.



The mass not fitting the linear curve e is

$E(\alpha,e) = P[y$ lies between $g_\alpha(x)$ and $e(x)]$.

Indeed this quantity constitutes a certain analogue of the excess mass where the line e plays the role of the level parameter λ. Again we assume a "support class" C, namely a class fulfilling

$\{x: e(x) < g_\alpha(x)\} \in C$ (e linear)

and thus modelling the expected deviations of e from g_α. In the present simple case C will be class of all intervals. How to estimate $E(\alpha,e)$? According to the procedure described

above we write $E(\alpha,\mathcal{C})$ as the supremum of an integral over $C \in \mathcal{C}$:

$$E(\alpha,\mathcal{C}) = \sup_{C \in \mathcal{C}} \mathbb{E}\ \mathrm{sign}(C)\ \mathrm{sign}_\alpha\{y > \mathcal{C}(x)\}$$

(Müller (1992)). Here sign_α denotes a the skew sign function defined as $\mathrm{sign}_\alpha(C) = (\mathrm{sign}(C) - 1)/2 + \alpha$. This quantity can be interpreted as the maximal correlation of the ("skew") residual pattern and the class $\mathcal{C}$. A plug-in of the empirical distribution leads to a natural estimator (Müller (1992)). This estimation method thus consists in minimizing the maximal correlation of the residual pattern with the class $\mathcal{C}$ (the "badness-of-fit method").

References.

Ehm, W. (1991), "Statistical problems with many parameters: critical quantities for approximate normality and posterior density based inference", Habilitationsschrift, Universität Heidelberg.

Good, I.J., and Gaskins, R.A. (1980), "Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data," *J.Amer.Statist. Assoc.* **75**, 42-73.

Groeneboom, P. (1985), "Estimating a monotone density," in *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II*, eds. L. LeCam, and R.Olshen, Wadsworth: Monterey, pp.539-555.

Hartigan, J.A. (1975), "Clustering Algorithms", Wiley: New York, London.

Hartigan, J.A. (1977), "Distribution problems in clustering," in *Classification and clustering*, ed. Van Ryzin, J., Academic Press: New York .

Hartigan, J.A. (1987), "Estimation of a convex density contour in two dimensions," *J. Amer. Statist. Assoc.* **82**, 267-270.

Hartigan, J.A. and Hartigan, P.M. (1985), "The DIP test of unimodality," *Ann. Statist.* **13**, 70 - 84.

Lientz B.P. (1970), "Results on nonparametric modal intervals", *SIAM J. Appl. Math.* **19**, 356-366.

Mammen, E. (1989), “Asymptotics with increasing dimension for robust regression with applications to the bootstrap”, *Ann. Statist.* **17**, 382-400 .

Mammen, E. (1992), “When does bootstrap work? Asymptotic Results and simulations”, *Lecture Notes in Statistics* **77**, Springer-Verlag: New York .

Martin-Löf, P. (1974), “The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data”, *Scand. J. Statist.* **1**, 3 -18.

Müller, D.W. (1980), “The analysis of ‘minimum guaranteed effect’ in the two sample case”, Preprint Nr. 72 (7/1980), Sonderforschungsbereich 123 "Stochastische Mathematische Modelle", Universität Heidelberg.

Müller, D.W. (1981), “Inference by means of total variation statistics”, in: Tagungsbericht 53/1981 of the Oberwolfach Conference on “Time Series and Density Estimation” (Dec.13-19, 1981) .

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

Müller, D.W. and Sawitzki, G. (1990) , “Using excess mass estimates to investigate the modality of a distribution”, in: Proceedings of the ICOSCO-I Conference (first International Conference on Statistical Computing, Çesme, Izmir 1987), Vol.II., American Science Press: Syracuse .

Müller, D.W. and Sawitzki, G. (1991), “Excess mass estimates and tests for multimodality”, *J.Amer. Statist. Assoc.* **86**, 738 - 746

Nolan, D. (1991), “The excess-mass ellipsoid”, *J.Multivariate Anal.***39**, 348-371.

Polonik, W. (1992), “The excess mass approach to cluster analysis and related estimation problems”, Preprint Nr.677, May 1992, Sonderforschungsbereich 123 “Stochastische Mathematische Modelle”, Universität Heidelberg.

Silverman, B.W. (1986), “Density estimation for statistics and data analysis”, Chapman and Hall: London.