Concepts of Outlyingness for Various Data Structures

Ursula Gather, Sonja Kuhnt and Jörg Pawlitschko

Department of Statistics, University of Dortmund

44221 Dortmund, Germany

Abstract. The term "outlier" is probably one of the vaguest and most imprecise ones in statistical science. There is no formal definition of an outlier, which all statisticians agree upon. However, for a univariate normal null-model Davies and Gather ([12] [13]) have introduced the concept of α-outliers and α-outlier regions, giving a definition which characterizes outliers only by their location relative to the assumed model for the good data. Outliers are thereby data points, observed in a region of the support of the anticipated distribution, namely an α-outlier region, where observations are – in a certain sense – unlikely under the assumed model. In this chapter we revisit this approach to outlyingness and generalize it to a variety of univariate and multivariate, continuous and discrete distributions as well as to structured models such as regression models and contingency tables. We also indicate how the concept of outlier regions can be used to define and construct outlier identification procedures.

Key words: Outliers, Tail regions, Density contours, Structured data, Robust procedures

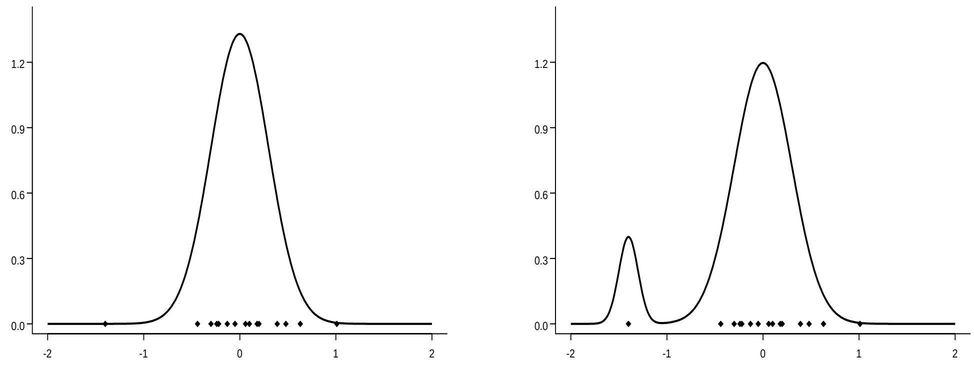
1 INTRODUCTION

In the statistical analysis of data we are often confronted with observations that "appear to be inconsistent with the remainder of that set of data" [2, p. 7] or, more generally, that "are far away [$\cdots$] from the pattern set by the majority of the data" [18, p. 25]. Observations of this kind are usually called "outliers". They may have a great impact on the statistical analysis and can cause completely misleading results when using standard methods. But also, sometimes outliers themselves provide the most interesting aspect of the data, for instance an unexpected long survival time in a clinical trial may indicate immunity against a certain disease. Although the problem of identifying and handling outliers has been subject of numerous investigations, there is no general agreement on a formal definition of outlyingness. Most authors, however, agree in that the term "outlier" is only meaningful if one has in mind a certain statistical model for the "good" data.

Consider the following set of data based on 15 observations made in 1846 on the vertical semi-diameter of the planet Venus. These observations are not the original measurements but the residuals with respect to a simple model as they have been analyzed by the astronomers Peirce and Chauvenet in 1852 and 1863, respectively (see [2, p. 38]):

-0.30	0.48	0.63	-0.22	0.18	-0.44	-0.24	-0.13
-0.05	0.39	1.01	0.06	-1.40	0.20	0.10	

At first glance it seems that the smallest (-1.40) and to a weaker extent the largest observation (1.01) are not consistent with the remaining values. They might therefore be "outliers". Figure 1 shows, besides the data itself, density functions of a symmetric, unimodal distribution and of a bimodal distribution. It is quite obvious that it will depend on the assumed (null-) distribution whether the observation -1.40 will be called an outlier or not.



$\mathcal{N}(0, 0.3)$ $\qquad$ $0.1\,\mathcal{N}(-1.4, 0.1) + 0.9\,\mathcal{N}(0.0, 0.3)$

Figure 1: Density functions for a normal and a mixture distribution

If one has in mind a normal distribution as on the left hand side of Figure 1, the smallest observation -1.40 clearly seems to be an outlier. But if one knew for some reason that the data generating mechanism must be described by a bimodal distribution as on the right hand side of Figure 1, then the value -1.40 is in accordance with the underlying model and will not be considered as outlying with respect to this null-distribution.

From another point of view, a mixture distribution as on the right hand side of Figure 1 can also be seen as a special case of an "outlier-generating model", where the outliers are, roughly spoken, assumed to be contaminated observations coming from a distribution that is different from the distribution of the "good" data. A comprehensive review of such models can be found e.g. in [2] or [15]. Such outlier-generating models have their merits. But they are based on a number of assumptions concerning the mechanism producing the outliers and thus impose too many restrictions. As a consequence, Davies and Gather (1993) proposed another way of formalizing the notion of outlyingness. Their approach tries to capture the element of surprise and of unlikeliness we associate with the term outlier. This is done by defining so-called outlier regions. Roughly spoken, these regions cover an area of the support of the anticipated null-distribution, where observations can only occur with a very small probability.

To fix ideas we may assume that in the above set of data the "good" part comes from a normal distribution $\mathcal{N}(0, 0.5)$. In Figure 2 that area under the density function which equals 0.1 has been marked, where the density function has smallest possible values. We denote all observations falling into the corresponding part of the support as 0.1-outliers. In our example these are the two values which we have already found suspicious.

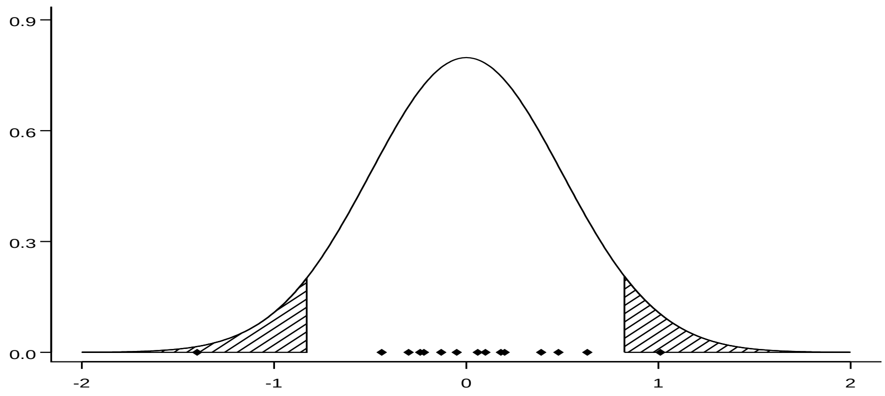


Figure 2: 0.1-outlier region of the $\mathcal{N}(0, 0.5)$-distribution

A formal definition of α-outliers for rather general statistical models is given in Section 2 below together with some remarks concerning procedures to identify such outliers in a given sample. In Section 3 we look at α-outlier regions for some univariate and multivariate continuous distributions. In the univariate case we investigate how these outlier regions are related to certain tail regions of the distribution. Section 4 contains a discussion of outlier regions in the continuous case when additional covariates are present. We focus on linear regression and one-way random effects models. Section 5 treats univariate and multivariate discrete distributions. For univariate discrete distributions with increasing-decreasing density function we present an algorithm for calculating the corresponding α-outlier regions. Section 6 covers structured discrete models like the logistic regression model and the loglinear Poisson model.

2 α-OUTLIER REGIONS

It is the aim to define α-outliers as objects which are in accordance with the idea that outliers are points which are rather unlikely under the true distribution. The concept of α-outliers has been introduced for univariate location-scale distributions in [12], [13] and has been extended to more general situations in [11]. The notion can be generalized to nearly all situations in which a clearly specified statistical model is assumed for the "good" data.

Let $\mathcal{P}$ be a family of distributions on a measurable space $(\mathcal{X}, \mathcal{A})$ which is dominated by a σ-finite measure ν such that $P \in \mathcal{P}$ has ν-density f. For $P \in \mathcal{P}$ let $\mathsf{supp}(P)$ denote the support of P and set $\mathsf{supp}(\mathcal{P}) = \bigcup_{P \in \mathcal{P}} \mathsf{supp}(P)$. For given $\alpha \in (0,1)$ the **α-outlier region of $P \in \mathcal{P}$** is then defined as

$$\mathsf{out}(\alpha, P) \;=\; \{x \in \mathsf{supp}(\mathcal{P})\colon\; f(x) < K(\alpha)\} \tag{1}$$

where

$$K(\alpha) \;=\; \sup\big\{K > 0\colon\; P\big(\{y\colon\; f(y) < K\}\big) \;\le\; \alpha\big\}.$$

With $\mathsf{inl}(\alpha, P) = \mathsf{supp}(\mathcal{P}) \setminus \mathsf{out}(\alpha, P)$ we define the corresponding **$\boldsymbol{\alpha}$-inlier region of $\boldsymbol{P}$**. Each point $x \in \mathsf{out}(\alpha, P)$ is called an α-outlier relative to P and each $x \in \mathsf{inl}(\alpha, P)$ an α-inlier.

For a random variable (r.v.) X with distribution P, one has by (1) that

$$P\big(X \in \mathsf{inl}(\alpha, P)\big) \ \geq \ 1 - \alpha, \tag{2}$$

and indeed, in most cases the α-inlier region of P can be viewed as the "smallest" subset of the support of $\mathsf{supp}(P)$ (with respect to the dominating measure ν) that has property (2).

The above definition of an α-outlier formalizes the perception that an outlier is a point which is extremely unlikely if P is the true distribution. Note that by (1) each $x \in \mathsf{supp}(\mathcal{P})$ and not only the observations of a given sample can be classified as outlying or inlying with respect to $P \in \mathcal{P}$. Note further that a r.v. X having distribution P may itself be observed in $\mathsf{out}(\alpha, P)$. However, because of (2) the probability of the corresponding event is smaller than α by definition.

When exploiting the definition of α-outlier regions for the task of outlier identification in a given sample, the sample size, say N, should be taken into account. If the sample points are assumed to be observations of i.i.d. random variables X_i, $i = 1, \ldots, N$, each with distribution P, then a natural choice of $\alpha = \alpha_N$ is given by

$$\alpha_N \ = \ 1 - (1 - \tilde{\alpha})^{1/N} \tag{3}$$

for some given $\tilde{\alpha} \in (0, 1)$. This choice guarantees that

$$P(X_i \in \mathsf{inl}(\alpha, P),\, i = 1, \ldots, N) \ \geq \ 1 - \tilde{\alpha}.$$

Under these assumptions the task of outlier identification can be formalized as follows: Given observations $\mathbf{x}_N = (x_1, \ldots, x_N)'$ with an unknown number of “good” data, which are assumed to come i.i.d. from a distribution P, find all those x_i, which are located in the outlier region $\mathsf{out}(\alpha_N, P)$ for an appropriately selected $\tilde{\alpha}$. Often $\tilde{\alpha}$ is chosen as 0.05 or 0.1.

Roughly spoken, there are two important types of outlier identification rules. A so-called simultaneous or one-step outlier identifier $OR(\alpha_N, \mathbf{x}_N) \subset \mathsf{supp}(\mathcal{P})$ can be viewed as an empirical version of $\mathsf{out}(\alpha_N, P)$. Each $x \in OR(\alpha_N, \mathbf{x}_N)$ is classified as α-outlier with respect to P. Since P is not known or only partially known, its unknown features have to be estimated from the data. The main problem here is that outliers in the data may seriously affect standard estimators of unknown distribution parameters, leading to the well-known effects of masking and swamping (cf. [13] for an in-depth discussion of this topic). This problem can be solved by using robust estimation methods. In [8] a one-step outlier identifier based on such robust estimators is called “resistant detection rule”.

Identification rules of the second type proceed stepwise by judging the outlyingness of the sample points in order of their “extremeness” relative to the sample. Inward testing procedures begin with the in some sense most suspicious observation. If this is declared as outlying, by some test for instance, it is removed from the sample and the procedure continues with the most extreme observation in the reduced sample. The procedure terminates if for the first time a subsample is found free of outliers or if a prespecified maximal number of possible outliers is reached. Outward testing procedures start with a reduced subsample that is supposed to contain no outliers. The least extreme of the observations not contained in this subsample is then checked with regard to its outlyingness. If it is indeed identified as an outlier, the procedure stops and only the observations in the reduced sample are considered as inlying. If not, it is joined with the subsample again and the least extreme of the remaining observations is checked. Since this paper is mainly concentrating on the concept of outlyingness in general rather than on identification rules, we will not give a further treatment of this topic here. For a deeper discussion see [16] and [26].

3 UNIVARIATE AND MULTIVARIATE CONTINUOUS DISTRIBUTIONS

In this section we investigate the shape of α-outlier regions when $(\mathcal{X}, \mathcal{P}) = (\mathbb{R}^d, \mathcal{B}^d)$, $d \in \mathbb{N}$, where $\mathcal{B}$ denotes the usual Borel-σ-algebra and $\mathcal{P} = \{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k\}$ is a family of distributions dominated by the d-dimensional Lebesgue-measure such that each $P_{\boldsymbol{\theta}} \in \mathcal{P}$ has density $f(\cdot, \boldsymbol{\theta})$.

A useful property of an α-outlier region in this context is got by the following lemma.

Lemma 1. For $P_{\boldsymbol{\theta}} \in \mathcal{P}$ assume that also $P^* \in \mathcal{P}$ where P^* is defined by $P^*(B) = P_{\boldsymbol{\theta}}(\mathbf{A}^{-1}(B - \boldsymbol{\mu}))$, $B \in \mathcal{B}^d$. Here $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\mathbf{A}$ denotes a regular $d \times d$-matrix. Then $\mathsf{out}(\alpha, P)$ is affine equivariant, i.e.

$$\mathsf{out}(\alpha, P^*) \;=\; \mathbf{A}\ \mathsf{out}(\alpha, P_{\boldsymbol{\theta}}) + \boldsymbol{\mu}.$$

Note however that $\mathsf{out}(\alpha, P_{\boldsymbol{\theta}})$ is not equivariant under more general transformations.

We first look at the case $d = 1$, that is we discuss outlier regions for univariate continuous distributions. As a first example we consider the univariate normal distribution. Taking the standard normal distribution and $\alpha = 0.1$,

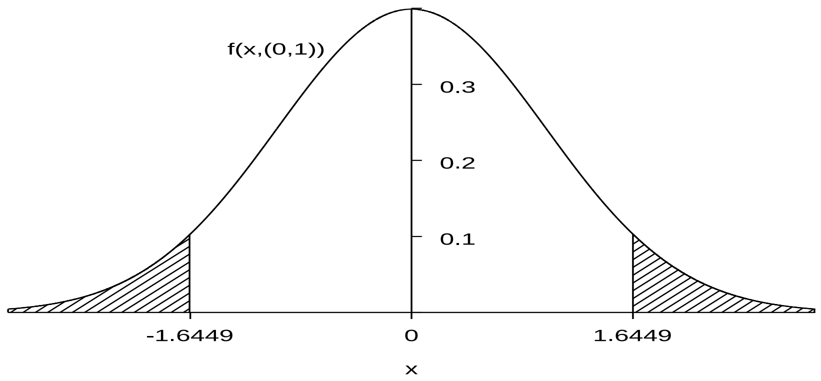


Figure 3: 0.1-outlier region of the standard normal distribution

the application of definition (1) gives

$$\mathsf{out}(0.1, \mathcal{N}(0,1)) \;=\; \big\{x\colon\; |x| > \; 1.6449\big\}.$$

This set (see Figure 3) is just the union of the upper and lower 0.05-tail region, see Lemma 2 below.

A close connection between tail and outlier regions holds in other cases as well. In the following, we make use of certain properties of a distribution $P_{\boldsymbol{\theta}}$. We call $P_{\boldsymbol{\theta}}$

(i) symmetric, if for some $\mu \in \mathbb{R}$ one has $f(\mu - x, \boldsymbol{\theta}) = f(\mu + x, \boldsymbol{\theta})$, $x \geq 0$,

(ii) having a strictly increasing-decreasing density, if for some $\mu_1, \mu_2 \in \mathbb{R}$ with $\mu_1 \leq \mu_2$ one has that $f(\cdot, \boldsymbol{\theta})$ is strictly increasing on $\mathsf{supp}(P_{\boldsymbol{\theta}}) \cap (-\infty, \mu_1]$, constant on $[\mu_1, \mu_2]$ and strictly decreasing on $\mathsf{supp}(P_{\boldsymbol{\theta}}) \cap [\mu_2, \infty)$,

(iii) having a strictly decreasing density, if $f(\cdot, \boldsymbol{\theta})$ is strictly decreasing on the entire support of $P_{\boldsymbol{\theta}}$.

If assumed, conditions (i)–(iii) need only to hold with probability one.

For $\alpha \in (0,1)$ let

$$q_\alpha(P_{\boldsymbol{\theta}}) \;=\; \inf\{x\colon\; P_{\boldsymbol{\theta}}(X \le x) \ge \alpha\} \tag{4}$$

denote the α-quantile of $P_{\boldsymbol{\theta}}$. Here X is random with distribution $P_{\boldsymbol{\theta}}$. Then $\mathsf{supp}(P_{\boldsymbol{\theta}}) \cap (-\infty, q_\alpha(P_{\boldsymbol{\theta}}))$ and $\mathsf{supp}(P_{\boldsymbol{\theta}}) \cap (q_{1-\alpha}(P_{\boldsymbol{\theta}}), \infty)$ define the lower and upper α-tail region of $P_{\boldsymbol{\theta}}$, respectively. Assume for simplicity that all distributions contained in $\mathcal{P}$ share the same support. Then the following simple relations hold.

Lemma 2.

(a) If $P_{\boldsymbol{\theta}}$ is symmetric, then $\mathsf{out}(\alpha, P_{\boldsymbol{\theta}})$ is equal to the union of the upper and lower $\alpha/2$-tail region of $P_{\boldsymbol{\theta}}$.

(b) If $P_{\boldsymbol{\theta}}$ has a strictly decreasing density, then $\mathsf{out}(\alpha, P_{\boldsymbol{\theta}})$ equals the upper α-tail region of $P_{\boldsymbol{\theta}}$.

Since in [13] the definition of $\mathsf{out}(\alpha, P_{\boldsymbol{\theta}})$ has been given only for the special case of the univariate normal distribution, part (a) of Lemma 2 has sometimes been mistaken as the actual definition. However, in many cases the two notions lead to different subsets of the support of $P_{\boldsymbol{\theta}}$. From Lemma 2, the α-outlier regions for many other important univariate distributions can readily be obtained. Examples are contained in Table 1.

Distribution	Parameter	Support	Lebesgue-density	α-outlier region
Normal	$\mu \in \mathbb{R},\ \sigma > 0$	$\mathbb{R}$	$\dfrac{1}{\sqrt{2\,\pi}\,\sigma}\ \exp\Big(-\dfrac{(x-\mu)^2}{2\,\sigma^2}\Big)$	$\big\{x:\ \lvert x-\mu\rvert > \sigma\, z_{1-\alpha/2}\big\}^{*}$
Cauchy	$\mu \in \mathbb{R},\ \sigma > 0$	$\mathbb{R}$	$\dfrac{1}{\pi\,\sigma}\ \dfrac{1}{1+\frac{(x-\mu)^2}{\sigma^2}}$	$\Big\{x:\ \lvert x-\mu\rvert > \sigma\ \tan\Big(\dfrac{\pi\,(1-\alpha)}{2}\Big)\Big\}$
Logistic	$\mu \in \mathbb{R},\ \sigma > 0$	$\mathbb{R}$	$\dfrac{\exp\Big(-\frac{x-\mu}{\sigma}\Big)}{\sigma\,\Big(1+\exp\Big(\frac{x-\mu}{\sigma}\Big)\Big)^2}$	$\big\{x:\ \lvert x-\mu\rvert > -\sigma\ \ln\big(1-\alpha/2\big)\big\}$
Double-Exponential	$\mu \in \mathbb{R},\ \sigma > 0$	$\mathbb{R}$	$\dfrac{1}{2\,\sigma}\ \exp\Big(-\dfrac{\lvert x-\mu\rvert}{\sigma}\Big)$	$\big\{x:\ \lvert x-\mu\rvert > -\sigma\ \ln\alpha\big\}$
Exponential	$\theta \in \mathbb{R},\ \lambda > 0$	$[\theta,\infty)$	$\dfrac{1}{\lambda}\ \exp\Big(-\dfrac{x-\theta}{\lambda}\Big)$	$\big\{x:\ x > \theta - \lambda\ \ln\alpha\big\}^{\dagger}$
Pareto	$\lambda > 0,\ \theta > 0$	$[\theta,\infty)$	$\dfrac{\lambda\,\theta^{\lambda}}{x^{\lambda+1}}$	$\big\{x:\ x > \theta\,\alpha^{-1/\lambda}\big\}^{\dagger}$

Table 1. α-outlier regions for some univariate continuous distributions

$*$ $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution

$\dagger$ if θ is fixed, else unite this set with $(-\infty,\theta)$

For univariate continuous distributions $P_{\boldsymbol{\theta}}$ which are not covered by Lemma 2, the relation of $\mathsf{out}(\alpha, P_{\boldsymbol{\theta}})$ to certain tail regions is more difficult. Sometimes, e.g. for mixture distributions, the α-outlier region may also contain an inner subset of the support. Although explicit expressions are seldom available, it is often possible to derive the boundary points of $\mathsf{out}(\alpha, P_{\boldsymbol{\theta}})$ numerically. For example, consider a distribution having a strictly increasing-decreasing density $f(\cdot, \boldsymbol{\theta})$ and let $F(\cdot, \boldsymbol{\theta})$ denote the corresponding distribution function. Suppose that $\lim_{x \searrow a_{P_{\boldsymbol{\theta}}}} f(x, \boldsymbol{\theta}) = \lim_{x \nearrow b_{P_{\boldsymbol{\theta}}}} f(x, \boldsymbol{\theta}) = 0$ where $a_{P_{\boldsymbol{\theta}}}$ and $b_{P_{\boldsymbol{\theta}}}$ denote the lower and upper bound of $\mathsf{supp}(P_{\boldsymbol{\theta}})$, respectively. Then the corresponding α-outlier regions can be obtained as follows:

Find points $x_1 < x_2$ within $\mathsf{supp}(P_{\boldsymbol{\theta}})$ such that the following two equations hold

$$\begin{aligned} \alpha &= 1 - F(x_2, \boldsymbol{\theta}) + F(x_1, \boldsymbol{\theta}), \\ f(x_1, \boldsymbol{\theta}) &= f(x_2, \boldsymbol{\theta}). \end{aligned} \tag{5}$$

The α-outlier region of $P_{\boldsymbol{\theta}}$ is then given as

$$\mathsf{out}(\alpha, P_{\boldsymbol{\theta}}) = \{x \in \mathsf{supp}(P_{\boldsymbol{\theta}})\colon\ x < x_1 \quad \text{or} \quad x > x_2\}.$$

Example 1. The (two-parameter) Weibull-distribution is an important parametric distribution with applications in reliability and lifetime analysis. It has Lebesgue-density

$$f(x, \beta, \lambda) = \lambda\, x^{\beta-1} \exp(-\lambda\, x^{\beta}), \qquad x \geq 0,$$

with shape parameter $\beta > 0$ and scale parameter $\lambda > 0$. The density $f(\cdot, \beta, \lambda)$ is strictly increasing-decreasing if and only if $\beta > 1$, otherwise it is strictly decreasing and Lemma 2 (b) can be applied. For $\beta > 1$, (5) can be written

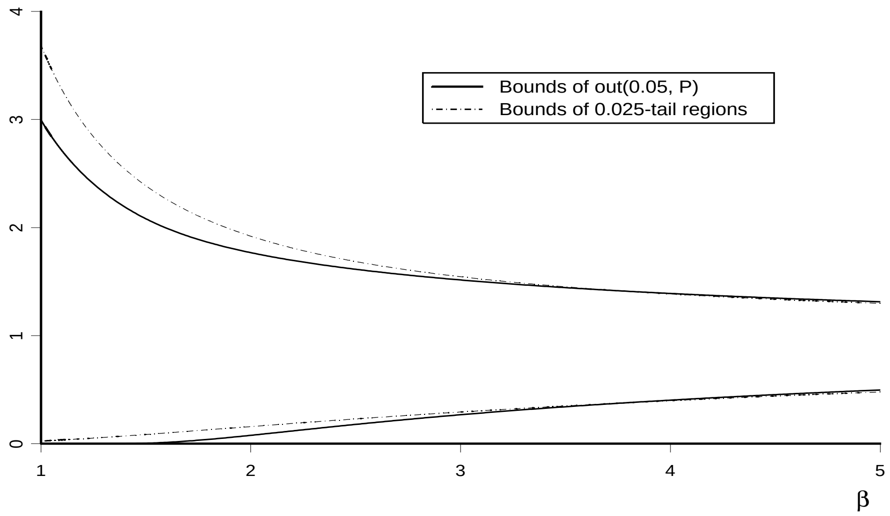


Figure 4: 0.05-outlier regions and upper and lower 0.025-tail regions for the Weibull-distribution with scale parameter $\lambda = 1$

as

$$\begin{aligned} 1 - \alpha - \exp(-\lambda\, x_1^{\beta}) + \exp(-\lambda\, x_2^{\beta}) &= 0, \\ x_1^{\beta-1}\, \exp(-\lambda\, x_1^{\beta}) - x_2^{\beta-1}\, \exp(-\lambda\, x_2^{\beta}) &= 0. \end{aligned}$$

These equations can readily be solved for x_1, x_2 by using e.g. the Newton-Raphson method.

Figure 4 displays the values of x_1 and x_2 as well as the $\alpha/2$- and $(1-\alpha/2)$-quantiles for various $\beta \geq 1$ in case that $\alpha = 0.05$ and $\lambda = 1$. Note that the α-inlier region is always smaller than the complement of the corresponding tail regions.

Example 2. As an example of a non-symmetric distribution with support equal to the entire real line we consider an extreme value distribution. It has

Lebesgue-density

$$f(x,\mu,\sigma) \;=\; \frac{1}{\sigma}\;\exp\Big(-\frac{x-\mu}{\sigma} - \exp\Big(-\frac{x-\mu}{\sigma}\Big)\Big), \qquad x \in \mathbb{R},$$

with location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma > 0$. From Lemma 1 it follows that it suffices to solve equations (5) for the standardized distribution with $\mu = 0$ and $\sigma = 1$. These equations reduce to

$$\begin{aligned} 1 - \alpha - \exp\big(-\exp(-x_2)\big) + \exp\big(-\exp(-x_1)\big) \;&=\; 0, \\ \exp(-x_1) - \exp(-x_2) + x_1 - x_2 \;&=\; 0. \end{aligned} \tag{6}$$

Take e.g. $\alpha = 0.05$. Solving (6) for x_1, x_2 yields the following 0.05-outlier region for the standardized extreme value distribution:

$$\mathsf{out}(0.05, P_{0,1}) \;=\; \{x \in \mathbb{R}\colon\; x < -1.5613 \quad \text{or} \quad x > 3.1615\}.$$

Note that the lower 0.025- and upper 0.025-tail region of $P_{0,1}$ are given by $(-\infty, -1.3053)$ and $(3.6762, \infty)$, respectively.

Outlier regions for a non-standardized extreme value distribution can be obtained by transforming the solutions of (6) according to $\tilde{x}_i = \sigma\, x_i + \mu$, $i = 1, 2$.

Within the framework of outlier regions, the problem of identifying outliers in univariate continuous distributions has thoroughly been investigated for the normal distribution and the (one-parameter) exponential distribution. The case of a normal distribution is treated in [13] and [16]. In [13], different types of identification rules are compared with respect to their worst-case behavior. It turns out that in case of unfavorably placed outliers identification rules based on robust estimators of location and scale still lead to satisfactory results. As an example we mention the one-step Hampel-identifier which is given by

$$OR^H(\mathbf{x}_N, \alpha_N) \;=\; \{x \in \mathbb{R}\colon\; |x - \mathsf{Med}(\mathbf{x}_N)| \;>\; g_N(\alpha_N)\; MAD(\mathbf{x}_N)\}.$$

Here $\mathsf{Med}(\mathbf{x}_N)$ denotes the sample median, $MAD(\mathbf{x}_N) = \mathsf{Med}(|x_i - \mathsf{Med}(\mathbf{x}_N)|)$ the (unstandardized) median absolute deviation from the median and $g_N(\alpha_N)$ an appropriately chosen normalizing constant. A common standardization is based on the requirement that under the null model H_0 - all observations come i.i.d. from a $\mathcal{N}(\mu, \sigma^2)$-distribution - one has

$$P_{H_0}\big(\text{no } X_i \text{ is identified as } \alpha_N\text{-outlier}\big) \;\geq\; 1 - \beta$$

for some appropriately chosen $\beta \in (0,1)$. Usually one takes $\beta = \tilde{\alpha}$, but other choices may be reasonable as well. In case of the Venus data from Section 1, application of the Hampel-identifier with $\beta = \tilde{\alpha} = 0.05$ yields $g_{15}(\alpha_{15}) = 6.36$ and $OR^H(\mathbf{x}_{15}, \alpha_{15}) = \mathbb{R}\backslash[-1.85, 1.97]$, hence no observation is flagged as outlying. The recommendations given in [13] are supported in [16] where also some considerations concerning the power of various identification rules are added. An interesting generalization of the Hampel-identifier for toxicological research is discussed in [31], where the case of non identically distributed random variables is treated whose expectations change according to a known toxicokinetical model.

The exponential distribution is investigated in [29] and [30]. The first paper is concerned with one-step identification rules whereas in [30] the focus lies on inward and outward testing procedures. As in the normal case, the use of robust estimators of the scale parameter when constructing identification rules is suggested. Best results concerning worst-case behavior and reasonable power are obtained with procedures based on a standardized sample median. Investigations for other univariate continuous distributions are rare, especially the case of non-symmetric distributions still awaits a complete treatment.

We now turn to multivariate extensions ($d \geq 2$). Whereas in the univariate continuous case α-outlier regions are often closely connected to certain tail regions, for multivariate continuous distributions the relation between the former and certain density contours seems helpful. This relation is especially apparent for a non-degenerate elliptically contoured distribution (see [14]). Such a distribution, denoted by $EC(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, has density with respect to the d-dimensional Lebesgue-measure given by

$$f(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \;=\; \frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\, g\big((\mathbf{x}-\boldsymbol{\mu})'\, \boldsymbol{\Sigma}^{-1}\, (\mathbf{x}-\boldsymbol{\mu})\big), \qquad \mathbf{x} \in \mathbb{R}^d. \tag{7}$$

Here $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Sigma}$ denotes a positive definite $d \times d$-matrix, and the function $g : \mathbb{R} \longrightarrow \mathbb{R}_+$ has to fulfill the condition

$$\int_0^\infty u^{d/2-1}\, g(u)\, du \;<\; \infty.$$

If in addition g is a strictly decreasing function on the positive real line, then

$$\mathsf{out}(\alpha, EC(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)) \;=\; \{\mathbf{x} \in \mathbb{R}^d : \; (\mathbf{x}-\boldsymbol{\mu})'\, \boldsymbol{\Sigma}^{-1}\, (\mathbf{x}-\boldsymbol{\mu}) > c_\alpha\}.$$

The constant c_α is determined by $P(\mathbf{X}'\mathbf{X} > c_\alpha) = \alpha$ where $\mathbf{X}$ denotes a random vector with distribution $EC(\mathbf{0}, \boldsymbol{I}_d, g)$. Hence, the α-outlier regions of an elliptically contoured distribution are complements of d-dimensional ellipsoids that are determined by certain density contours. The special case of a non-degenerate multivariate normal distribution corresponds to choosing

$$g(u) \;=\; \frac{1}{(2\pi)^{d/2}}\, \exp(-u).$$

In this case, the well known results for quadratic forms of independent normally distributed random variables yield $c_\alpha = \chi^2_{d,1-\alpha}$, the $(1-\alpha)$-quantile of the chi-square distribution with d degrees of freedom.

The identification of outliers in the multivariate normal case using one-step identification rules has been investigated in [4] and [5]. The focus lies on

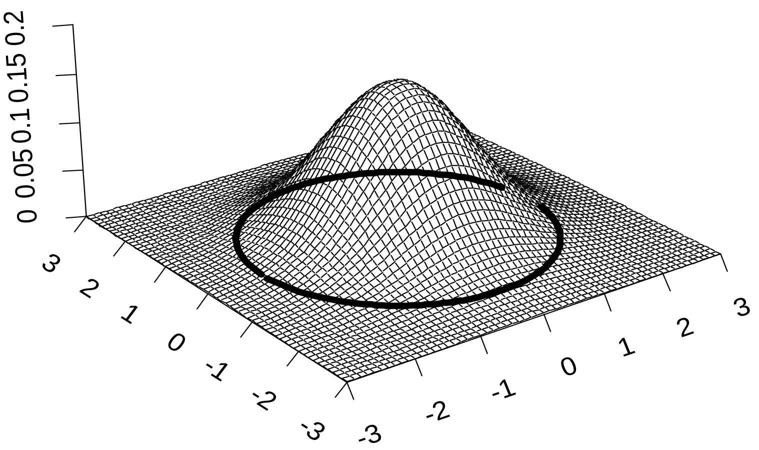


Figure 5: Density contour of a bivariate standard normal distribution

rules of the type

$$
\begin{aligned}
&OR(\mathbf{x}_N, \alpha_n) \\
&= \{\mathbf{x} \in \mathbb{R}^d : (\mathbf{x} - \mathbf{m}_N(\mathbf{x}_N))' \, \mathbf{S}_N(\mathbf{x}_N)^{-1} \, (\mathbf{x} - \mathbf{m}_N(\mathbf{x}_N)) > g_N(\alpha_N)\},
\end{aligned}
$$

where $\mathbf{S}_N(\mathbf{x}_N)$ denotes a robust estimator of the covariance matrix and $\mathbf{m}_N(\mathbf{x}_N)$ a corresponding robust estimator of $\boldsymbol{\mu}$. For example one may consider the so-called minimum volume ellipsoid which is the smallest ellipsoid containing at least $[(N+d-1)/2]$ data points. The mean and sample covariance matrix of the corresponding subsample yield highly robust estimators of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$. Stepwise rules for outlier identification in the multivariate normal case have been proposed e.g. by [9]. However, a discussion of those rules in the light of outlier regions is still wanting.

An interesting application to the problem of outlier identification in online monitoring data from intensive care medicine can be found in [3]. Another useful application can be found in [19], where α-outlier regions of a multivariate normal distribution are investigated as a tool in cluster analysis. More generally, in [19] a fix point cluster is defined as a subset of a given data set such that a certain simultaneous outlier identifier (constructed for a pre-specified cluster reference distribution) detects no α-outlier in this subset. In [19], [20] this approach is transferred to cluster analysis for linear regression (cf. Section 4) and discrete data.

For other multivariate continuous distributions explicit representations for α-outlier regions can seldom be given. A simple case where this is still possible is given in the following example.

Example 3. Consider the case of a bivariate exponential distribution when the marginal distributions are independent. That is, we consider a bivariate Lebesgue-density

$$f(x, y, \lambda_1, \lambda_2) \;=\; \frac{1}{\lambda_1\,\lambda_2}\,\exp\Bigl(-\frac{x}{\lambda_1}-\frac{y}{\lambda_2}\Bigr), \qquad x, y > 0,$$

with scale parameters $\lambda_1, \lambda_2 > 0$.

Again, α-outlier regions for this bivariate distribution can be found by using their relation to certain density contours which are given as certain straight lines. From (1) one finds that here an α-outlier region is given by

$$\mathsf{out}(\alpha, P_{\lambda_1,\lambda_2}) \;=\; \Bigl\{(x, y)' \in \mathbb{R}_+ \times \mathbb{R}_+ :\; y > \lambda_2\,\Bigl(-\ln\bigl(\lambda_1\,\lambda_2\,K\bigr) - \frac{x}{\lambda_1}\Bigr)\Bigr\}$$

where K is determined as solution of the equation

$$\lambda_1\,\lambda_2\,K\,\Bigl(-\ln\bigl(\lambda_1\,\lambda_2\,K\bigr)+1\Bigr) \;=\; \alpha.$$

Hence, the α-outlier region of this bivariate exponential distribution is the complement (in the first quadrant of $\mathbb{R}^2$) of a simplex with vertices $(0,0)'$, $(0, -\lambda_1\,\ln(\lambda_1\,\lambda_2\,K))'$, and $(-\lambda_2\,\ln(\lambda_1\,\lambda_2\,K), 0)'$. For other bivariate extensions of the exponential distribution that also allow for dependence, the construction of the corresponding α-outlier regions is more complicated.

4 STRUCTURED CONTINUOUS DATA

So far we have investigated outlier regions for data situations where a homogeneous population was assumed. We focus now on the case of inhomogeneous populations, especially in the presence of certain covariates.

We first consider the simple linear regression model:

$$Y = \beta_0 + \mathbf{X}' \boldsymbol{\beta}_1 + U \tag{8}$$

where Y denotes the response, $\mathbf{X} \in \mathbb{R}^p$ a vector of regressors, and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1')' \in \mathbb{R}^{p+1}$ a vector of regression parameters. With $U = Y - (\beta_0 + \mathbf{X}' \boldsymbol{\beta}_1)$ we denote the corresponding residual. Suppose that a random sample of independent copies $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, N$, from model (8) is given. To exploit the concept of α-outlier regions successfully in this regression set-up, we have to make appropriate assumptions for the distribution of the response and the joint distribution of the regressors if they are random. Reasonable assumptions are

$$P_{Y|\mathbf{X}} = \mathcal{N}(\beta_0 + \mathbf{X}' \boldsymbol{\beta}_1, \sigma^2), \tag{R1}$$

for a scale parameter $\sigma^2 > 0$, that is the conditional distribution of Y given the vector of regressors is normal, and

$$P_{\mathbf{X}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{R2}$$

that is the joint distribution of the regressors is a p-variate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and $(p \times p)$-covariance matrix $\boldsymbol{\Sigma}$. Note that under

(R1) the conditional distribution of the residual U is given by

$$P_{U|\mathbf{X}} \;=\; \mathcal{N}(0,\sigma^2).$$

Further, if (R1) and (R2) hold, then

$$P_{(Y,\mathbf{X})} \;=\; \mathcal{N}\left(\begin{pmatrix}\beta_0 + \boldsymbol{\mu}'\,\boldsymbol{\beta}_1 \\ \boldsymbol{\mu}\end{pmatrix}, \begin{bmatrix}\sigma^2 + \boldsymbol{\beta}_1'\,\boldsymbol{\Sigma}\,\boldsymbol{\beta}_1 & \boldsymbol{\beta}_1'\,\boldsymbol{\Sigma} \\ \boldsymbol{\Sigma}\,\boldsymbol{\beta}_1 & \boldsymbol{\Sigma}\end{bmatrix}\right). \tag{9}$$

There are several reasonable ways of defining outlier regions for this regression set-up. First look at the case where we only assume (R1). Then according to the general definition (1), a response-α-outlier region could be defined as

$$\mathsf{out}(\alpha, P_{Y|\mathbf{X}}) \;=\; \big\{y \in \mathbb{R}\colon\;\; u = |y - (\beta_0 + \mathbf{X}'\,\boldsymbol{\beta}_1)| > \sigma\; z_{1-\alpha/2}\big\}. \tag{10}$$

This type of outlier region is especially useful in the case of fixed regressors where only outliers in y-direction are of interest. Note that essentially the residual u determines whether some y is outlying or not, a large absolute value of u indicates an α-outlier.

However, often also the regressors are random quantities as in many econometric or sociometric applications. In this case, also outliers in $\mathbf{x}$-direction are of interest. Under assumption (R2), a regressor-α-outlier region can be defined as in Section 2 for the multivariate normal distribution:

$$\mathsf{out}(\alpha, P_{\mathbf{X}}) \;=\; \{\mathbf{x} \in \mathbb{R}^p\colon\; (\mathbf{x} - \boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x} - \boldsymbol{\mu}) > \chi^2_{p,1-\alpha}\}. \tag{11}$$

This approach leads to a population based formalization of the concept of so-called leverage points which in a given sample $(Y_i, \mathbf{X}_i)$ are “cases for which $\mathbf{X}_i$ is far away from the bulk of the $\mathbf{X}_i$ in the data” [28].

From (9), outliers in y- as well as in $\mathbf{x}$-direction can now be characterized by the α-outlier region

$$\begin{aligned}&\mathsf{out}(\alpha, P_{(Y,\mathbf{x})}) \ = \\ &\left\{(y, \mathbf{x}')' \in \mathbb{R}^{p+1} \colon \ \frac{(y - (\beta_0 + \mathbf{x}'\,\boldsymbol{\beta}_1))^2}{\sigma^2} + (\mathbf{x} - \boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x} - \boldsymbol{\mu}) > \chi^2_{p+1,1-\alpha}\right\}.\end{aligned}$$

This is again the α-outlier region of a certain multivariate normal distribution, where the special structure of the mean vector and covariance matrix is taken into account.

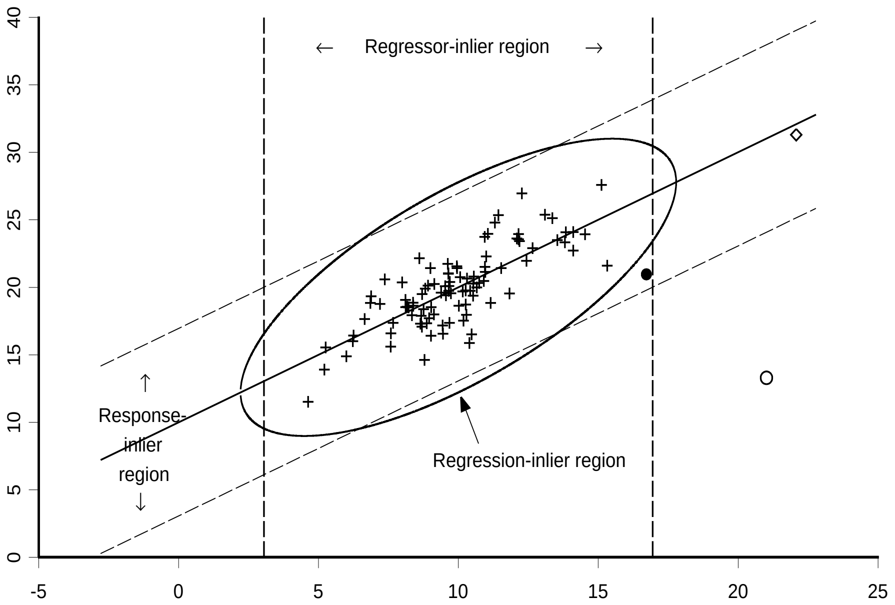


Figure 6: Outlier regions in the regression set-up

In Figure 6 the three types of outlier regions are shown for the case that in (R1) and (R2) we have $N = 100$, $\mu = 10$, $\beta_0 = 10$, $\beta_1 = 1$, $\sigma^2 = 1$ and $\boldsymbol{\Sigma}$ equals 4. Further, $\alpha = \alpha_N = 1 - (1 - \tilde{\alpha})^{1/N}$ with $\tilde{\alpha} = 0.05$. The filled circle "$\bullet$" indicates a point which is a regression-outlier but neither a regressor- or response-outlier, the diamond "$\diamond$" indicates a "good", the unfilled circle "$\circ$" a "bad" leverage point.

Several outlier regions for linear regression models have also been investigated in [6]. However, with exception of (10) these regions do not correspond to the outlier regions presented here. For an unspecified measurable function $g : \mathbb{R}^{p+1} \longrightarrow \mathbb{R}$, depending on α and the model parameters, in [6] a more general (g, α)-outlier region is defined in terms of

$$\mathsf{out}(g, \alpha, P_{(Y,\mathbf{X})}) \;=\; \{(y, \mathbf{x}')' \in \mathbb{R}^{p+1} : \; g(y, \mathbf{x}) > 1\}.$$

By judicious choice of g this more general definition also covers other notions of outlyingness. Note however that the interpretation of an α-outlier region as given in Section 2 does no longer hold true for a general (g, α)-outlier region.

We investigate two functions g that yield reasonable (g, α)-outlier regions: Choosing

$$g(y, \mathbf{x}) \;=\; \frac{|y - \beta_0 - \mathbf{x}'\boldsymbol{\beta}_1|}{\sigma\, c_\alpha \sqrt{1 + (\mathbf{x} - \boldsymbol{\mu})'\, \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}} \tag{12}$$

leads to a (g, α)-outlier region by which leverage points are only classified as outlying if the corresponding residual has a quite large absolute value. On the other hand, choosing

$$g(y, \mathbf{x}) \;=\; \frac{|y - \beta_0 - \mathbf{x}'\boldsymbol{\beta}_1|}{\sigma\, c_\alpha} \sqrt{1 + (\mathbf{x} - \boldsymbol{\mu})'\, \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \tag{13}$$

leads to a (g, α)-outlier region that will classify only "good" leverage points as inlying. In [6] these regions are called outlier regions of Type III and Type II, respectively (outlier regions of Type I correspond to response-outlier regions as defined in (10)). The constant c_α has to be chosen in accordance with an appropriate standardization. Figure 7 contains both regions in case that the relevant parameters are chosen as in Figure 6. The normalizing constant has been chosen such that the probability of the occurrence of a (g, α_N)-outlier in a sample of size $N = 100$ does not exceed $\alpha_N = 1 - (1 - \tilde{\alpha})^{1/N}$ for

$\tilde{\alpha} = 0.05$. This requirement leads to $c_{\alpha_N} = 3.13$ for a Type III-outlier region and $c_{\alpha_N} = 6.83$ for a Type II-outlier region, where these values are obtained by simulations.

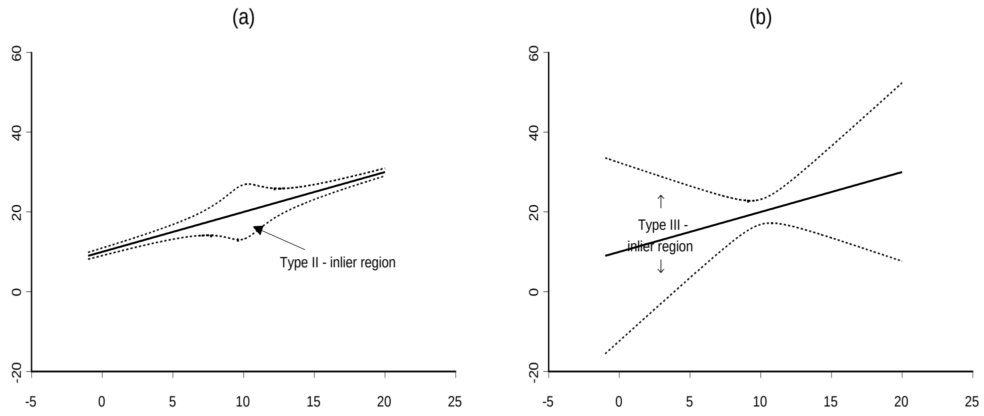


Figure 7: Outlier regions of (a) Type II and (b) Type III in the regression set-up

In [6] it is shown that identification of outliers in the regression set-up based on empirical versions of response- and Type II-outlier regions leads inevitably to procedures with bad worst-case behavior. In general, procedures based on least median of squares or least trimmed squares as estimators of the regression parameters (cf. [27]) yield the most satisfactory results.

Outlier regions for ANOVA models with fixed effects can be defined as in (10). For a two-way ANOVA, one-step-identification rules implicitly based on this definition of outlyingness are discussed in [33]. If random effects must be included the situation becomes more complex. Consider the most simple case of a one-way random effects model with k classes and N_i measurements taken in each class, $i = 1, \ldots, k$. Typical applications of this model include the statistical analysis of interlaboratory studies. The definition of outlier regions for this set-up has been investigated in [34]. Let Y_{ij} denote the j-th outcome in the i-th class, then we assume

$$Y_{ij} \;=\; \mu + U_i + E_{ij}, \qquad i = 1, \ldots, k,\ j = 1, \ldots, N_i,$$

where $\mu \in \mathbb{R}$ is the overall mean, U_i is the unobservable random effect of the i-th class and E_{ij} represents the measurement error for the j-th observation in this class. Set $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iN_i})'$. We make the additional assumption

$$P_{(\mathbf{Y}_i, U_i)} \;=\; \left(\begin{pmatrix} \mathbf{1}_{N_i}\,\mu \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_e^2\,\mathbf{I}_{N_i} + \sigma_u^2\,\mathbf{1}_{N_i}\,\mathbf{1}'_{N_i} & \mathbf{1}_{N_i}\,\sigma_u^2 \\ \mathbf{1}'_{N_i}\,\sigma_u^2 & \sigma_u^2 \end{bmatrix} \right), \tag{RE}$$

where $\mathbf{I}_m$ denotes the $(m \times m)$-identity matrix, $\mathbf{1}_m \in \mathbb{R}^m$ is the vector having all components equal to one, and σ_u^2, $\sigma_e^2 > 0$ are the so-called variance components representing the contributions of the random effect and the measurement error to the total variance of the Y_{ij}. We assume further that the random vectors $(\mathbf{Y}_i', U_i)'$, $i = 1, \ldots, k$, are independent. Note that from (RE) the conditional distribution of Y_{ij} given the random effect U_i equals

$$P_{Y_{ij}|U_i} \;=\; \mathcal{N}(\mu + U_i, \sigma_e^2),$$

and the marginal distributions of the joint observations in a single class and of the random effects are respectively given as

$$\begin{aligned} P_{\mathbf{Y}_i} \;&=\; \mathcal{N}\big(\mathbf{1}_{N_i}\,\mu, \sigma_e^2\,\mathbf{I}_{N_i} + \sigma_u^2\,\mathbf{1}_{N_i}\,\mathbf{1}'_{N_i}\big), \\ P_{U_i} \;&=\; \mathcal{N}(0, \sigma_u^2). \end{aligned} \tag{14}$$

With the exception of $P_{\mathbf{Y}_i}$ these distributions do not depend on the indices i and j. For $P_{\mathbf{Y}_i}$, dependence occurs only with regard to the sample size. There are now different forms of suspicious observations which should be treated separately. Firstly one may be interested in a location-α-outlier within the i-th class. The corresponding α-outlier region is simply given by

$$\mathsf{out}(\alpha, P_{Y_{ij}|U_i}) \;=\; \{y \in \mathbb{R}\colon\ |y - \mu - u_i| > \sigma_e\, z_{1-\alpha/2}\},$$

where the level could be chosen as $\alpha = \alpha_{N_i} = 1 - (1 - \tilde{\alpha})^{1/N_i}$ or $\alpha = \alpha_N = 1 - (1 - \tilde{\alpha})^{1/N}$ with $N = \sum_{i=1}^{k} N_i$.

Besides the observations it may be the classes themselves which are of interest. E.g., in an interlaboratory study one may observe a laboratory that in general has conspicuous results compared with the other labs. From (14) a class-α-outlier region could be defined as

$$\begin{aligned}\mathsf{out}(\alpha, P_{\mathbf{Y}_i}) = \Big\{ & \mathbf{y} = (y_{i1}, \ldots, y_{iN_i})' \in \mathbb{R}^{N_i} : \\ & \frac{\sum_{j=1}^{N_i} (y_{ij} - \mu)^2}{\sigma_e^2} - \frac{\sigma_u^2}{\sigma_e^2} \, \frac{\left(\sum_{j=1}^{N_i} y_{ij} - N_i \, \mu \right)^2}{\sigma_e^2 + N_i \, \sigma_u^2} > \chi^2_{1-\alpha, N_i} \Big\},\end{aligned}$$

which, similarly to $\mathsf{out}(\alpha, P_{(Y,\mathbf{X})})$ in the regression model, is the α-outlier region of a multivariate normal distribution with a certain mean and covariance structure.

In [34] it is mentioned that one may be interested in different types of outlying classes: such which seem to differ markedly from the others with respect to location and such which show a comparably larger (or smaller) dispersion of the outcomes. The first type of aberrant classes can be described by a location-α-outlier region within the random effects defined by

$$\mathsf{out}(\alpha, P_{U_i}) = \{ u \in \mathbb{R}\colon \ |u| > \sigma_u \, z_{1-\alpha/2} \}.$$

Note that this is just the α-outlier region of a univariate normal distribution with variance σ_u^2. For the second type, in [34] an outlier region is suggested that is based on a scale estimator of σ_e. Let $s_m : \mathbb{R}^m \longrightarrow \mathbb{R}_+$ denote such an estimator which operates on a set of m observations and which is location and scale equivariant. Then in [34] a scale-α-outlier region with respect to s_{N_i} is defined by

$$\mathsf{out}(\alpha, P_{\mathbf{Y}_i}, s_{N_i}) = \{ \mathbf{y} \in \mathbb{R}^{N_i} : \ \left| \ln(s_{N_i}(\mathbf{y})) - \ln(\sigma_e) \right| > c_\alpha \}.$$

The constant c_α is determined from

$$P\big(\mathbf{Y}_i \in \mathsf{out}(\alpha, P_{\mathbf{Y}_i}, s_{N_i})\big) = \alpha.$$

This scale-α-outlier region is an α-outlier region in the sense of definition (1) only if the distribution of $\ln(s(\mathbf{Y}_i))$ is indeed symmetric with center $\ln(\sigma_e)$. For the outlier regions defined for the classes it is convenient to choose $\alpha = \alpha_k = 1 - (1 - \tilde{\alpha})^{1/k}$.

A great difference between outlier regions for the one-way random effects model and the other models discussed so far is that some of them depend on the unobservable U_i's. For practical applications, e.g. when constructing a simultaneous outlier identifier, the realizations of the random effects must be predicted from the data. As is shown in [34], reliable outlier identifiers can be constructed with robust estimators of the model parameters μ, σ_u, σ_e, and robust predictors of the random effects. Especially procedures based on medians yield results that are quite satisfactory.

5 UNIVARIATE AND MULTIVARIATE DISCRETE DISTRIBUTIONS

The α-outlier concept can also successfully be applied to discrete distributions which have an enumerable support. Hence in contrast to Section 3 families of distributions $\mathcal{P} = \{P_{\boldsymbol\theta}, \boldsymbol\theta \in \Theta \subset \mathbb{R}^k\}$ on $(\mathbb{R}^d, \mathcal{B}^d)$ are considered which are dominated by the counting measure on $\mathsf{supp}(P_{\boldsymbol\theta})$. Let the discrete density of $P_{\boldsymbol\theta}$ be denoted by $p(\cdot, \boldsymbol\theta)$ and the α-quantile of $P_{\boldsymbol\theta}$ be defined as in (4). The lower and upper α-tail regions of $P_{\boldsymbol\theta}$ are again given by $\mathsf{supp}(P_{\boldsymbol\theta}) \cap (-\infty, q_\alpha(P_{\boldsymbol\theta}))$ and by $\mathsf{supp}(P_{\boldsymbol\theta}) \cap (q_{1-\alpha}(P_{\boldsymbol\theta}), \infty)$, respectively. Note that the upper α-tail region is empty if $q_{1-\alpha}(P_{\boldsymbol\theta}) = \max(\mathsf{supp}(P_{\boldsymbol\theta}))$, a similar result holds for the lower tail region.

Usually, the α-outlier regions with respect to discrete distributions cannot be obtained with similar arguments as those for continuous distributions (as for example in Table 1). Lemma 1 still holds but is not useful in most cases as usually $P^* \notin \mathcal{P}$. A somewhat artificial example where $P^* \in \mathcal{P}$ is always true is given by the general family of discrete uniform distributions $\mathcal{U} = \{U(n,a,h),\ n \in \mathbb{N},\ a \in \mathbb{R},\ h \in \mathbb{R}\}$, where $U(n,a,h)$ has density

$$p(x,a,n,h) \;=\; \frac{1}{n+1}\,\mathbf{1}_{\{a\,i+h,\,i\in\{0,\dots,n\}\}}(x),$$

that is $\mathsf{supp}(U(n,a,h)) = \{x \in \mathbb{R}\colon x = a + i\,h,\ i \in \{0,\dots,n\}\}$ and $\mathsf{supp}(\mathcal{U}) = \mathbb{R}$ (see [22, p. 272]). Following the definition of an α-outlier region given in Section 1, the α-outlier region of $U(n,a,h)$ is given by

$$\mathsf{out}(U(n,a,h)) \;=\; \mathbb{R} \setminus \mathsf{supp}(U(n,a,h)) \;=\; a\,\mathsf{out}(U(n,1,0)) + h.$$

Hence, irrespectively of the choice of $\alpha \in (0,1)$, each $x \notin \mathsf{supp}(U(n,a,h))$ is an α-outlier with respect to $U(n,a,h)$.

The same is true for Lemma 2, which can be formulated for the discrete case as well. But it is not helpful in most cases. Take as an example the family of binomial distributions $Bin(n,\pi)$ with fixed positive integer $n \in \mathbb{N}$ and parameter $\pi \in [0,1]$, where the distributions have support $\{0,\dots,n\}$ and density function

$$p(x,\pi) \;=\; \binom{n}{x}\,\pi^x\,(1-\pi)^{n-x}\,\mathbf{1}_{\{0,\dots,n\}}(x).$$

Each binomial distribution has a strictly increasing-decreasing density with $\mu_1 = \mu_2 = [\![\pi(n+1)]\!]$ if $\pi(n+1) \notin \mathbb{N}$, and $\mu_1 = \pi(n+1) - 1$, $\mu_2 = \pi(n+1)$ otherwise. If $\pi = 0.5$, all $Bin(n,0.5)$ are symmetric around $\mu = \frac{n}{2}$, $n \in \mathbb{N}$. In this case Lemma 2 (a) can be applied and the α-outlier region is the union of the corresponding upper and lower $\alpha/2$-tail regions of $Bin(n, 0.5)$. These can easily be found using standard tables. If for example $n = 5$ the lower 0.05-tail region equals $\{0\}$ and the upper 0.05-tail region equals $\{5\}$. Following Lemma 2 (a) the 0.1-outlier region of $Bin(5, 0.5)$ therefore contains the values 0 and 5. For $\pi \neq 0.5$, the α-outlier region of $Bin(n, \pi)$ can be derived by calculating all probabilities and applying the definition. Especially with higher values of n this becomes rather tedious. Figure 8 shows the case $n = 6$, $\pi = 0.6$.

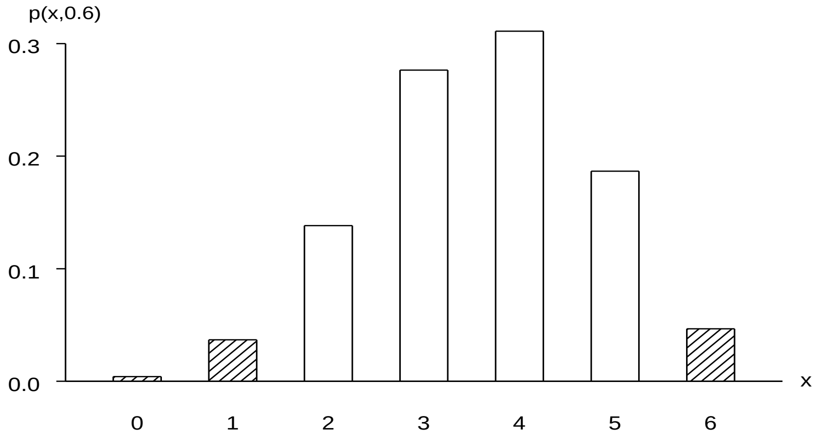


Figure 8: 0.1-outlier region of the binomial distribution $Bin(6, 0.6)$

Consider the following condition. We say $P_{\boldsymbol{\theta}}$

(iv) has an increasing-decreasing density if for some $\mu \in \mathsf{supp}(P_{\boldsymbol{\theta}})$ one has $p(x_1, \boldsymbol{\theta}) \leq p(x_2, \boldsymbol{\theta})$ for all $x_1 \leq x_2 \leq \mu$ and $p(x_1, \boldsymbol{\theta}) \geq p(x_2, \boldsymbol{\theta})$ for all $\mu \leq x_1 \leq x_2$, $x_1, x_2 \in \mathsf{supp}(P_{\boldsymbol{\theta}})$.

Condition (iv) is assumed to hold with probability one. It is more general than those given in Section 3 as it is fulfilled by all distributions for which one condition out of (i) to (iii) holds.

Inlier regions for a distribution $P_{\boldsymbol{\theta}}$ fulfilling condition (iv) can be derived using the following procedure. Let α be given. For $x_q \in \mathsf{supp}(P_{\boldsymbol{\theta}})$ denote $\inf\{(x \in \mathsf{supp}(P_{\boldsymbol{\theta}})\colon x > x_q\}$ by x_{q+1} and $\sup\{x \in \mathsf{supp}(P_{\boldsymbol{\theta}})\colon x < x_q\}$ by x_{q-1}.

1. Set $L = U = 0$, $x_0 = \mu$, $x_{sup} = \sup(\mathsf{supp}(P_{\boldsymbol{\theta}}))$, $x_{inf} = \inf(\mathsf{supp}(P_{\boldsymbol{\theta}}))$.

2. Set $L = L - i$ where $i \in \mathbb{N}$, $p(x_{L-i}, \boldsymbol{\theta}) = p(x_L, \boldsymbol{\theta})$ and $p(x_{L-i-1}, \boldsymbol{\theta}) \neq p(x_L, \boldsymbol{\theta})$. If $x_U = x_{sup}$ go to step 4, else continue.

3. Set $U = U + j$ where $j \in \mathbb{N}$, $p(x_{U+j}, \boldsymbol{\theta}) = p(x_U, \boldsymbol{\theta})$ and $p(x_{U+j+1}, \boldsymbol{\theta}) \neq p(x_U, \boldsymbol{\theta})$.

4. If $P(X \in [x_L, x_U]) \geq 1 - \alpha$ then $\mathsf{inl}(\alpha, P) = [x_L, x_U] \cap \mathsf{supp}(P_{\boldsymbol{\theta}})$, else
if $x_U = x_{sup}$ set $L = L - 1$ and return to step 2, else
if $x_L = x_{inf}$ set $U = U + 1$ and return to step 3, else
set $U = U + 1$ if $p(x_{U+1}) = \max(p(x_{U+1}, \boldsymbol{\theta}), p(x_{L-1}, \boldsymbol{\theta}))$ and $L = L - 1$ if $p(x_{L-1}, \boldsymbol{\theta}) = \max(p(x_{U+1}, \boldsymbol{\theta}), p(x_{L-1}, \boldsymbol{\theta}))$ and return to step 2.

Procedures like this have already been applied to the binomial and the Poisson distribution in [10] and [22]. As an example consider the family of Poisson distributions $Poi(\lambda)$, $\lambda \in \mathbb{R}_+$, with density

$$p(x, \lambda) \;=\; \frac{\lambda^x}{x!}\, \exp(-\lambda)\, \mathbf{1}_{\mathbb{N}}(x)$$

and support $\mathsf{supp}(Poi(\lambda)) = \mathbb{N}$. Since $p(\cdot, \lambda)$ is increasing-decreasing, the algorithm presented above can be applied. For some values of α and λ the corresponding outlier regions are given in Table 2.

Note that although the definition of an α-outlier provides unambiguous outlier regions for a given distribution, these can be the same for different values of α and different parameters.

	Parameter λ		
α	3	3.5	4
0.01	$\mathbb{N} \setminus \{0, \ldots, 8\}$	$\mathbb{N} \setminus \{0, \ldots, 8\}$	$\mathbb{N} \setminus \{0, \ldots, 9\}$
0.05	$\mathbb{N} \setminus \{0, \ldots, 6\}$	$\mathbb{N} \setminus \{0, \ldots, 7\}$	$\mathbb{N} \setminus \{1, \ldots, 8\}$
0.1	$\mathbb{N} \setminus \{1, \ldots, 6\}$	$\mathbb{N} \setminus \{1, \ldots, 6\}$	$\mathbb{N} \setminus \{1, \ldots, 7\}$

Table 2: α-outlier regions for some Poisson distributions

Outlier identification procedures for samples from binomial and Poisson distributions can be derived from procedures for the logistic regression model and the loglinear Poisson model ([22], [10]). The definition of outliers for these more complex data structures will be discussed in Section 6.

We consider next the case of multivariate discrete distributions. In most cases the marginal distributions are equal to well known univariate distributions. But it is in general not possible to find a representation of the outlier region in terms of the outlier regions of the marginal distributions. Even in the case of independence where the joint distribution equals the product of the marginal distributions $\mathsf{out}\big(\alpha, \bigotimes_{i=1}^{d} P_i\big) = \times_{i=1}^{d} \mathsf{out}(\alpha, P_i)$ does not necessarily hold. A similar result has been noted in [11].

As an example consider the case of a two-dimensional random vector $(X_1, X_2)'$ and the following alternative distributions: The so-called bivariate Poisson distribution (see [21], Chapter 37.2) has bivariate density

$$\begin{aligned} & p(x_1, x_2, \lambda_1, \lambda_2, \lambda_{12}) \\ = \; & \exp(-(\lambda_1 + \lambda_2 + \lambda_{12})) \sum_{i=0}^{\min(x_1, x_2)} \frac{\lambda_1^{x_1 - i} \, \lambda_2^{x_2 - i} \, \lambda_{12}^{i}}{(x_1 - i)! \, (x_2 - i)! \, i!} \; \mathbf{1}_{\mathbb{N}}(x_1) \; \mathbf{1}_{\mathbb{N}}(x_2), \end{aligned}$$

with parameters λ_1, λ_2, $\lambda_{12} \in \mathbb{R}_+$. In contrast to this, the multiple Poisson distribution given by $Poi(\lambda_1 + \lambda_{12}) \otimes Poi(\lambda_2 + \lambda_{12})$ has bivariate density

$$\begin{aligned} & p(x_1, x_2, \lambda_1, \lambda_2, \lambda_{12}) \\ = \; & \exp(-(\lambda_1 + \lambda_2 + 2\,\lambda_{12}))\, \frac{(\lambda_1 + \lambda_{12})^{x_1}\, (\lambda_2 + \lambda_{12})^{x_2}}{x_1!\, x_2!}\, \mathbf{1}_{\mathbb{N}}(x_1)\, \mathbf{1}_{\mathbb{N}}(x_2), \end{aligned}$$

with the same parameters. In both cases the marginal distributions are Poisson distributions with parameters $\lambda_1 + \lambda_{12}$ and $\lambda_2 + \lambda_{12}$, respectively. Corresponding 0.1-inlier regions of two multivariate distributions with parameter values $\lambda_1 = 1$, $\lambda_2 = 3$ and $\lambda_{12} = 0.5$ are given in Figure 9 as well as the product of the 0.1-inlier regions of the marginal distributions. Apparently these regions are different and there is no obvious connection between the inlier regions and hence nor between the outlier regions of the marginal and the joint distributions.

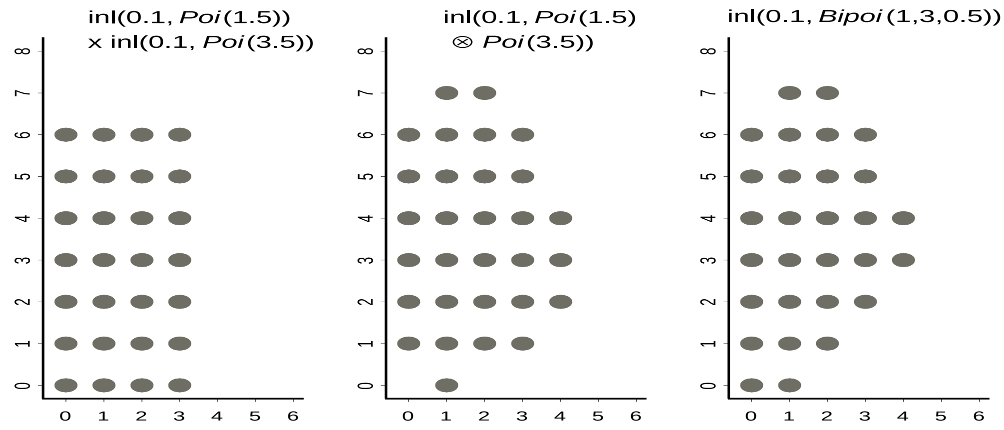


Figure 9: 0.1-inlier regions of some joint and marginal Poisson distributions

The inlier regions given in Figure 9 have been derived by lengthy calculations of density values and the application of definition (1). So far no algorithms or outlier identification procedures have been developed for multivariate discrete distributions. The main focus has been on structured situations which will be discussed in the next section.

6 STRUCTURED DISCRETE DATA

As in the continuous case, discrete data are often sampled from an inhomogeneous population such that the assumption of identical distributions for all observations is not valid . However, quite often there exists an underlying structure connecting the individual distributions. This is taken into account by using a model like the logistic regression model, the loglinear Poisson model, or the Poisson regression model. All these models may be embedded into the larger class of generalized linear models (see [24]).

We deal here with samples of a fixed number N of discrete random variables $Y_1, \ldots, Y_N$ which are independently, but not identically distributed. Their distributions belong to the same family $\mathcal{P} = \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ but are characterized by different parameter values $\boldsymbol{\theta}_i = \boldsymbol{\theta}(\mathbf{x}_i, \boldsymbol{\beta})$, depending on some fixed covariates $\mathbf{x}_i \in \mathbb{R}^p$ and a common parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$.

Take as an example the loglinear poisson model for a contingency table where the vector $\mathbf{Y} = (Y_1, \ldots, Y_N)'$ contains the frequency counts for the cells of the table. Using the representation as generalized linear model, each loglinear poisson model can be described by a fixed $p \times N$-design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, containing only the entries $-1, 0, 1$, and a parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$ (see [1, p. 80] and [24, p. 13]). The cell counts Y_i are assumed to be independent and to follow Poisson distributions $Poi(\lambda_i)$ with parameters $\lambda_i = \lambda(\mathbf{x}_i, \boldsymbol{\beta}) = \exp(\mathbf{x}_i' \boldsymbol{\beta})$, $i = 1, \ldots, N$.

Following Section 5 an α-outlier region $\mathsf{out}(\alpha, \otimes_{i=1}^N P_{\boldsymbol{\theta}(\mathbf{x}_i, \boldsymbol{\beta})})$ for the joint distribution of the Y_i can be considered. Note that in this case each vector $(y_1, \ldots, y_N) \in \mathsf{supp}(\otimes_{i=1}^N P_{\boldsymbol{\theta}(\mathbf{x}_i, \boldsymbol{\beta})})$ is either classified as outlier or inlier. In the contingency table set-up this would be the complete table. Such an application of the α-outlier concept is useful for a sample of more than one table describing the same context. This is the case when we have outcomes of different experiments on the same topic with e.g. the final goal of estimating a common odds ratio.

Usually, however, only one contingency table is observed and one is interested rather in single cell counts, hence one needs outlier regions for the distributions of the individual Y_i, $i = 1, \ldots, N$. The same applies for other models than the loglinear poisson model. We therefore look at outlier regions $\mathsf{out}(\alpha_i, P_{\boldsymbol{\theta}_i})$ of the marginal distributions. As in the case of i.i.d. random variables the values α_i should be chosen such that the probability of the occurrence of an outlier in the whole sample does not exceed a given $\tilde{\alpha}$ (see (3)). Again, this can be achieved by setting $\alpha_i = 1 - (1 - \tilde{\alpha})^{1/N}$ for all $i = 1, \ldots, N$.

Example 4. Consider the loglinear independence model for a 3×3 table, where $N = 9$ and $p = 5$. The parameters of the marginal Poisson distributions are given by $\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ with design matrix

$$\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_9}) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \\ 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 \end{bmatrix}.$$

As parameter vector we choose $\boldsymbol{\beta} = (4, 0.2, -0.2, 0.4, 0.3)'$. Table 3 contains the 0.01-outlier regions for the nine cell distributions. For example, a count of 85 observed in cell 3 would be classified as 0.01-outlier. However, the same count observed in cell 1 would be an 0.01-inlier due to the inhomogeneous cell distributions.

cell	λ_i	$\mathsf{out}(0.01, Poi(\lambda_i))$
1	99.48	$\mathbb{N} \setminus \{75, \ldots, 126\}$
2	90.02	$\mathbb{N} \setminus \{67, \ldots, 115\}$
3	33.12	$\mathbb{N} \setminus \{19, \ldots, 48\}$
4	66.69	$\mathbb{N} \setminus \{47, \ldots, 88\}$
5	60.34	$\mathbb{N} \setminus \{41, \ldots, 80\}$
6	22.20	$\mathbb{N} \setminus \{11, \ldots, 35\}$
7	81.45	$\mathbb{N} \setminus \{59, \ldots, 105\}$
8	73.70	$\mathbb{N} \setminus \{52, \ldots, 96\}$
9	27.11	$\mathbb{N} \setminus \{15, \ldots, 41\}$

Table 3: Loglinear Poisson model: 0.01-outlier regions for Example 4

In the loglinear Poisson model the design matrix $\mathbf{X}$ is always fixed, whereas in the logistic regression and the Poisson regression model stochastic regressors X_i, $i = 1, \ldots, N$, are possible as well. In both cases it is assumed that conditioned on the regressors the Y_i are independent but not identically distributed with binomial distributions

$$P_{Y_i|\mathbf{X}_i} \;=\; Bin\Big(m_i, \frac{1}{1+\exp(-\mathbf{X}_i'\boldsymbol{\beta})}\Big) \tag{15}$$

in case of the logistic regression model, and Poisson distributions

$$P_{Y_i|\mathbf{X}_i} \;=\; Poi\big(m_i \, \exp(-\mathbf{X}_i' \boldsymbol{\beta})\big) \tag{16}$$

in case of the Poisson regression model respectively. Here $m_1, \ldots, m_N$ are given positive integers.

As in the situation of the linear regression model discussed in Section 4, the concept of α-outlier regions can be applied in two different ways: The first way consists in specifying the joint distributions of $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, N$, and then deriving the corresponding α_i-outlier regions. Since unlike to the continuous case there is no natural choice for these joint distributions, usually the conditional distributions (15) and (16) are considered. Following this approach yields again to the same α_i-outlier regions as obtained with fixed explanatory variables (see [11]).

The concept of outlier regions for structured discrete data and corresponding outlier identification methods have recently been considered by a few authors, see [10], [11], [17], [22]. Again, outlier identification procedures that are based on robust estimators of the unknown parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$ give satisfactory results. The main problem here consists in finding such an estimator.

y_{ij}	$j=1$	2	3	4	5
$i=1$	18	41	41	20	21
2	39	20	20	22	22
3	24	20	20	16	18
4	20	20	19	19	19
5	23	19	20	17	20

$\widehat{\lambda}_{ij}$	1	2	3	4	5
1	24.2	21.0	21.0	20.0	21.0
2	25.3	22.0	22.0	20.9	22.0
3	23.0	20.0	20.0	19.0	20.0
4	21.9	19.0	19.0	18.1	19.0
5	23.0	20.0	20.0	19.0	20.0

Table 4: Invented 5×5-table and median polish estimates

Example 5. As an example of the identification of outliers in structured discrete data we consider a data set which was invented by [32]. These data are also discussed in [2, p. 438] and are displayed in Table 4 in their original form as a 5×5-table. We have chosen this example, because here, unlike to the usual situation in contingency tables, the outlying cells are rather obvious, namely given by the observations y_{11}, y_{13}, and y_{21}. As robust estimation method in the case of the independence model for two-way tables an application of the median polish method to the logarithm of the table has been suggested in [25]. The resulting estimates $\widehat{\lambda}_{ij}$ are given in the right hand side of Table 4. The outlier regions $\mathsf{out}(\alpha, Poi(\widehat{\lambda}_{ij}))$ based on these estimators can now be viewed as estimates of the true outlier regions $\mathsf{out}(\alpha, Poi(\lambda_{ij}))$. This approach leads quite naturally to the procedure of identifying all observations lying in these estimated outlier regions as outliers.

	$j=1$	2	3	4	5
$i=1$	$\{12,\ldots,37\}$	$\{10,\ldots,33\}$	$\{10,\ldots,33\}$	$\{10,\ldots,32\}$	$\{10,\ldots,33\}$
2	$\{13,\ldots,38\}$	$\{11,\ldots,34\}$	$\{11,\ldots,34\}$	$\{10,\ldots,33\}$	$\{11,\ldots,34\}$
3	$\{12,\ldots,36\}$	$\{10,\ldots,32\}$	$\{10,\ldots,32\}$	$\{9,\ldots,31\}$	$\{10,\ldots,32\}$
4	$\{11,\ldots,34\}$	$\{9,\ldots,31\}$	$\{9,\ldots,31\}$	$\{8,\ldots,29\}$	$\{9,\ldots,31\}$
5	$\{12,\ldots,36\}$	$\{10,\ldots,32\}$	$\{10,\ldots,32\}$	$\{9,\ldots,31\}$	$\{10,\ldots,32\}$

Table 5: Inlier regions $\mathbb{N} \setminus \mathsf{out}(0.01, Poi(\widehat{\lambda}_{ij}))$ in Example 5

Table 5 contains the 0.01-inlier regions with respect to the Poisson distributions given by the median polish estimates. According to the proposed method an observation not located in the corresponding estimated inlier region is identified as outlier. This is the case for the observations y_{11}, y_{13} and y_{21}, hence the correct result is obtained. This example used a simplified version of special one-step outlier identification procedures for contingency tables which have been introduced in [22], [17], and [23].

Acknowledgement

The financial support of the Deutsche Forschungsgesellschaft (SFB 475) is gratefully acknowledged.

References

[1] Agresti, A. *Categorical Data Analysis*, Wiley, New York (1990).

[2] Barnett, V. and Lewis, T. *Outliers in Statistical Data*, 3^{rd} ed., Wiley, New York (1994).

[3] Bauer, M., Gather, U. and Imhoff, M. The identification of multiple outliers in online monitoring data. Technical Report 29/1999, SFB 475, University of Dortmund (1999).

[4] Becker, C. and Gather, U. The masking breakdown point of multivariate outlier identification rules. *J. Am. Statist. Ass.* **94**, 947–955 (1999).

[5] Becker, C. and Gather, U. The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules. *Comput. Statist. Data Anal.* **36**, 119–127 (2001).

[6] Boscher, H. *Behandlung von Ausreißern in linearen Regressionsmodellen.* Thesis, Department of Statistics, University of Dortmund (1992).

[7] Brown, M.B. Identification of the source of significance in two-way contingency tables. *Appl. Statist.* **23**, 405–413 (1974).

[8] Carey, V.J., Walters, E.E., Wager, C.G. and Rosner, B.A. Resistant and test-based outlier rejection: Effects on Gaussian one- and two-sample inference. *Technometrics* **39**, 320–330 (1997).

[9] Caroni, C. and Prescott, P. Sequential Application of Wilks's Multivariate Outlier Test. *Applied Statistics* **41**, 355–364 (1992).

[10] Christmann, A. *Ausreißeridentifikation und robuste Schätzer im logistischen Regressionsmodell.* Thesis, Department of Statistics, University of Dortmund (1992).

[11] Christmann, A. *On positive breakdown point estimators in regression models with discrete response variables.* Habilitationsschrift, Department of Statistics, University of Dortmund (1998).

[12] Davies, L. and Gather, U. The identification of multiple outliers. Technical Report 89/1, Department of Statistics, University of Dortmund (1989).

[13] Davies, L. and Gather, U. The identification of multiple outliers. *J. Am. Statist. Ass.* **88**, 782–792 (1993).

[14] Fang, K.-T., Kotz, S. and Ng, K.-W. *Symmetric Multivariate and Related Distributions.* Chapman and Hall, London (1990).

[15] Gather, U. Modelling the occurence of multiple outliers. *Allg. Statist. Archiv* **74**, 413–428 (1990).

[16] Gather, U. and Becker, C. Outlier identification and robust methods. In *Handbook of Statistics, Vol. 15: Robust Methods* (Edited by G.S. Maddala and C.R. Rao), pp. 123–143, Elsevier, Amsterdam (1997).

[17] Gather, U., Becker, C. and Kuhnt, S. Robust methods for complex data structures. In *Data Analysis, Classification and Related Methods.* (Edited by H.A.L. Kiers et al.), pp. 315–320, Springer, Berlin (2000).

[18] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. *Robust Statistics – The Approach Based on Influence Functions.* Wiley, New York (1986).

[19] Hennig, C. Clustering and outlier identification: Fixed point cluster cnalysis. In *Advances in Data Science and Classification* (Edited by A. Rizzi et al.), pp. 37–42, Springer, Berlin (1998).

[20] Hennig, C. Models and methods for clusterwise linear regression. In *Classification in the Information Age* (Edited by W. Gaul and H. Locarek-Junge), pp. 179–187, Springer, Berlin (1999).

[21] Johnson, N.L., Kotz, S. and Balakrishnan, N. *Discrete Multivariate Distributions*. Wiley, New York (1997).

[22] Kuhnt, S. *Ausreißeridentifikation im Loglinearen Poissonmodell für Kontingenztafeln unter Einbeziehung robuster Schätzer*. Thesis, Department of Statistics, University of Dortmund (2000).

[23] Kuhnt, S. Outliers in contingency tables. In *Proceedings of the 6th International Conference on Computer Data Analysis and Modeling*, Minsk, Belarus (2001).

[24] McCullagh, P. and Nelder, J.A. *Generalized Linear Models*. 2^{nd} ed., Chapman and Hall, London (1989).

[25] Mosteller, F. and Parunak, A. Identifying extreme cells in a sizable contingency table: probabilistic and exploratory approaches. In *Exploring Data Tables, Trends and Shapes* (Edited by D.C. Hoaglin et al.) pp. 189–224, Wiley, New York (1985).

[26] Rosner, B. On the detection of many outliers. *Technometrics* **17**, 221–227 (1975).

[27] Rousseeuw, P.J. and Leroy, A.M. *Robust Regression and Outlier Detection*, Wiley, New York (1987).

[28] Rousseeuw, P.J. and van Zoomeren, B.C. Unmasking multivariate outliers and leverage points. *J. Am. Statist. Ass.* **85**, 633–639 (1990).

[29] Schultze, V. and Pawlitschko, J. Identification of outliers in exponential samples with stepwise procedures. Technical Report 56/2000, SFB 475, University of Dortmund (2000).

[30] Schultze, V. and Pawlitschko, J. The identification of outliers in exponential samples. *Statist. Neerlandica*, to appear.

[31] Selinski, S. and Becker, C. Outlier detection in experimental data using a modified Hampel identifier. Technical Report 40/2001, SFB 475, University of Dortmund (2001).

[32] Simonoff, J.S. Detecting outlying cells in two-way contingency tables via backwards-stepping. *Technometrics* **30**, 339–345 (1988).

[33] Terbeck, W. and Davies, L. Interactions and outliers in the two-way analysis of variance. *Ann. Statist.* **26**, 1279–1305 (1998).

[34] Wellmann, J. and Gather, U. Identification of outliers in a one-way random effects model. *Statist. Papers*, to appear.