Simultaneous small sample inference based on profile likelihood

Von der Naturwissenschaftlichen Fakultät
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades eines

Doktors der Gartenbauwissenschaften

- Dr. rer. hort. -

genehmigte Dissertation
von

Dipl.-Ing. agr. Daniel Gerhard

geboren am 22.03.1979, in Castrop-Rauxel

2010

Referent:	Prof. Dr. L. A. Hothorn
Korreferent:	Prof. Dr. D. Hauschke
Tag der Promotion:	09.06.2010

Abstract

Generalized linear models allow the parameter estimation under assumption of a distribution of the exponential family to summarize collected data. Based on the parameters of those models, tests can be conducted to reject a Null-hypothesis, or confidence intervals can be calculated, giving an idea about the uncertainty of the parameter location with a given error rate. If multiple parameters are of interest, multiplicity adjustment has to be performed to control a global error rate for all tests or confidence intervals.

In this thesis multiple tests and simultaneous confidence intervals are constructed based on deviance profiles, which show directly the uncertainty about the location of a parameter conditional on the given data. Statistics are obtained by minimizing the deviance conditional on a parameter of interest, allowing for a linear transformation by a prespecified contrast matrix by considering additional constraints. Under assumption of an approximately multivariate normal distribution for these statistics, multiple tests and simultaneous confidence intervals are constructed. As the correlation structure of the Normal distribution is unknown, it is directly estimated from the data.

In contrast to Wald-type confidence intervals, which are more simple to calculate, the profile based intervals allow for unequal distances of the lower and upper confidence limits to the point estimates; this might provide better coverage probability under assumption of a non-equitailed distribution. A further advantage of the profile based intervals is their transformation invariance.

The validity of the discussed methods is illustrated by means of a simulation study focusing on count and categorical data. Part of this thesis is a user-friendly implementation of the methods in the statistical software package R, which is presented by an evaluation of several small case studies.

Keywords: profile likelihood, simultaneous confidence intervals, multiple comparisons

Zusammenfassung

Generalisierte lineare Modelle bieten die Möglichkeit, Parameter unter Annahme einer Verteilung aus der Exponentialfamilie zu schätzen, um die erhobenen Daten adequat zusammenzufassen. Basierend auf diesem Modell können Tests zum Verwerfen einer Nullhypothese, oder Konfidenzintervalle zur Abschätzung der Unsicherheit über die Lokation eines Parameters zu einer vorgegebenen Fehlerrate berechnet werden. Bei mehreren Parametern von Interesse ist eine Multiplizitätsadjustierung erforderlich, um eine globale Fehlerrate simultan für alle Tests oder Konfidenzintervalle einzuhalten.

Die vorgelegte Arbeit befasst sich mit der Konstruktion von multiplen Tests und simultanen Konfidenzintervallen basierend auf Devianz-Profilen, die direkt die Unsicherheit über die Lokation eines Parameters konditional der Daten widerspiegeln. Teststatistiken erhält man, indem konditional auf dem jeweiligen Parameter von Interesse die Devianz minimiert wird, wobei zusätzliche Bedingungen auch die Transformation der Parameter über eine zuvor gewählte Kontrastmatrix erlauben. Unter der Annahme, daß diese Statistiken approximativ einer multivariaten Normalverteilung folgen, können multiple Tests und simultane Konfidenzintervalle berechnet werden. Da die Korrelationsstruktur der multivariaten Normalverteilung jedoch unbekannt ist, wird diese direkt aus den Daten geschätzt.

Im Gegensatz zu den einfacher zu berechnenden Wald-Konfidenzintervallen erlauben die Profil-basierten Intervalle eine unterschiedliche Distanz der oberen und unteren Konfidenzgrenze zum Punktschätzer, was zu einer besseren Einhaltung der Überdeckungswahrscheinlichkeit bei Annahme nicht normalverteilter Daten führen kann. Ein weiterer Vorteil ist die Transformationsinvarianz der Profil-basierten Intervalle.

Die Validität der vorgestellten Methoden wird über eine umfangreiche Simulationsstudie mit dem Schwerpunkt auf Zähl- und kategorialen Daten dargestellt. Ein Teil der Arbeit umfasst die nutzerfreundliche Implementierung der Methoden in die statistische Software R, welche anhand einer Reihe von Fallbeispielen vorgestellt wird.

Schlagworte: Likelihood Profile, Simultane Konfidenzintervalle, Multiple Vergleiche

Contents

List of abbreviations

CI Confidence Interval

GLM Generalized Linear Model

HOA Higher Order Approximations

MLE Maximum Likelihood Estimate

QA Quadratic Approximation

SRDP Signed Root Deviance Profile

SRDP 0 modified Signed Root Deviance Profile excluding complete zero counts

General notation

$\boldsymbol{y} = (y_i)$ Vector of observations with $i = 1, \ldots, N$

$\boldsymbol{X} = (y_{ij})$ Design matrix of covariates with $i = 1, \ldots, N$, and $j = 1, \ldots, k$

$\boldsymbol{\beta} = (\beta_j)$ Parameter vector in a generalized linear model with $j = 1, \ldots, k$

$\boldsymbol{\Sigma}_j = (\sigma_{jj})$ Variance-covariance matrix of $\hat{\beta}_j$ with $j = 1, \ldots, k$

$\boldsymbol{\epsilon} = (\epsilon_i)$ Vector of residuals with $i = 1, \ldots, N$

$\eta = (\eta_i)$ Linear predictor in a generalized linear model

$\mu = (\mu_i)$ Vector of predictions on the original scale

$\tau(\cdot)$ Link function

$L(\cdot), l(\cdot), D(\cdot)$ Likelihood, log-likelihood, and deviance functions

$d(\cdot)$ Deviance statistic

$r(\cdot)$ Signed root deviance statistic

$r^*(\cdot)$ Modified signed root deviance statistic

$\tilde{r}(\cdot)$ Scaled signed root deviance statistic

$t(\cdot)$ Wald statistic

$j(\cdot)$ Fisher information function

$\boldsymbol{C} = (c_{mj})$ Contrast matrix with $j = 1, \ldots, k$, and $m = 1, \ldots, M$

$\boldsymbol{\psi} = (\psi_m)$ $m = 1, \ldots, M$ linear combinations of model parameters

$\boldsymbol{\Sigma}_\psi = \left(\sigma_{mm}^{(\hat{\psi})}\right)$ Variance-covariance matrix of $\hat{\psi}_m$ with $m = 1, \ldots, M$

α type-I-error rate

$q_{1-\alpha}$, $z_{1-\alpha}$ Quantile of a χ^2, or Normal-distribution

δ_l, δ_u Lower and upper confidence limits

ϕ Dispersion parameter

Chapter 1

Introduction

A common method to summarize observed data is to calculate arithmetic means and standard deviations. To evaluate the uncertainty about these means, confidence intervals can be constructed by adding and subtracting a specific constant times the standard error around the mean values. This kind of data summary demands an equal-tailed distribution of these parameters, which is naturally fulfilled for normal distributed data. In many applications the observed data are not continuous measurements, but counts, or categorical outcomes. These values are located on a restricted space; counts being only positive integers, proportions are limited between 0 and 1. The more the mean value approaches these limits of the parameter space, the less appropriate is the assumption of an equal-tailed distribution. Hence, the summarization of the data by mean and standard deviation becomes more inadequate and requires a more difficult interpretation. This problem becomes evident, when, for example rare events are counted; the arithmetic mean is located near zero and subtracting two times the standard deviation will result in a negative lower limit, which is certainly inappropriate to characterize the variability of the average of several counts.

Assuming an adequate skewed distribution for a non-normally distributed response of interest is of course more appropriate than using the normality assumption. Instead of estimating the mean and the standard deviation separately, a dependence of the variance on the mean has to be assumed, making it necessary to estimate both parameters simultaneously. This can be done by evaluating a corresponding likelihood, representing the probability of the location of a parameter given the observed data. Estimate of means

are calculated by finding the maximum of this likelihood function, whereas the standard error for the mean is a measure for the likelihood curvature, showing the uncertainty about the location of this maximum. But the likelihood function offers more information about the parameter of interest than summarized by mean and standard error, which might be preferable to use when constructing confidence intervals. By calculating several likelihood statistics around the location of the maximum likelihood estimate, a likelihood profile can be established. Confidence limits are then estimated by searching for the likelihood statistics that correspond to a quantile of the underlying distribution of the parameter of interest. This leads to confidence intervals, which are most strongly supported by the data, as they closely incorporate the curvature of the likelihood function.

The problem of non-normally distributed parameters is mostly covered in the area of nonlinear regression. For diagnosis of nonlinear models assuming a Gaussian distributed response, Bates and Watts (1988) introduced the profile-t-plot, visualizing the deviation from a normal distribution for single parameters of interest. Additionally, likelihood based confidence intervals can be obtained by their method. Chen and Jennrich (1996) proposed a generalization of these profiles to any maximum likelihood estimation. They suggest the profiling of arbitrary functions of parameters, based on the likelihood transformation invariance, showing the wide range of applications for this method, which is discussed for applied nonlinear regression by Quinn et al. (1999). Confidence intervals based on their likelihood profiles are direct inversions of a likelihood ratio test, which in many cases might be preferable over linear test statistics (Donaldson and Schnabel, 1987). The profiles are constructed equivalently by optimization or spline interpolation for several t- or z-statistics conditional on some fixed parameters around the MLE (Quinn et al., 2000). An alternative method constructing profiles by integration is suggested by Venzon and Moolgavkar (1988), providing similar results.

The likelihood profiles are only based on a first order approximation, whereas inappropriate results may be obtained for small samples. Therefore, adjustments of the likelihood by higher order approximations have become popular, producing nearly exact results even at small sample sizes. For categorical data for example, Staicu (2009) shows eminent bias reduction by applying higher order approximations for confidence intervals of a binomial parameter or a log odds ratio. There is already a vast amount of literature available on the field of higher order approximations, most of it is not directly aimed at using these

methods in practice; but there are also nice contributions, giving some understanding for their application, e.g. Brazzale et al. (2007), Brazzale and Davison (2008), Skovgaard (2001), and especially Pierce and Peters (1992).

Although the advantages of likelihood based methods are widely known, they are not frequently used. Reasons might be longer and more expensive computations and fewer implementations into statistical software packages.

So far, issues about simultaneous inference are not discussed on the basis of profile likelihood confidence intervals or hypotheses testing. There are several publications available for example on simultaneous confidence intervals comparing several binomial proportions; for example considering Wald intervals for the difference of proportions with Sheffé and Bonferroni-type intervals in Piegorsch (1991), or utilizing a multivariate normal distribution with an estimated correlation structure in Schaarschmidt et al. (2008). Bonferroni-adjusted simultaneous confidence intervals based on score statistics are presented in Agresti et al. (2008).

In this thesis, the adaptation of multiple contrast tests and simultaneous confidence intervals to existing profile likelihood methods is described with a restriction to generalized linear models with quite simple parameter layouts. Focus is set on the applicability of this likelihood based approach to real data problems, evaluating the performance with respect to alternative methods. Furthermore, the additional information obtained from displaying deviance profiles is discussed. At last an implementation as package in the statistical software R (R Development Core Team, 2009) is presented to calculate simultaneous profile deviance confidence intervals and multiple tests for user-defined contrasts.

Chapter 2

Example data

2.1 Angina: Dose response study

The first example is a dataset from a dose response study of a drug to treat Angina pectoris, taken from Westfall Peter H. (1999). Response variable was the duration of pain-free walking after treatment, relative to the values before treatment. Four dose steps and a negative control were applied with ten observations for each treatment level. The data is shown in Figure 2.1.

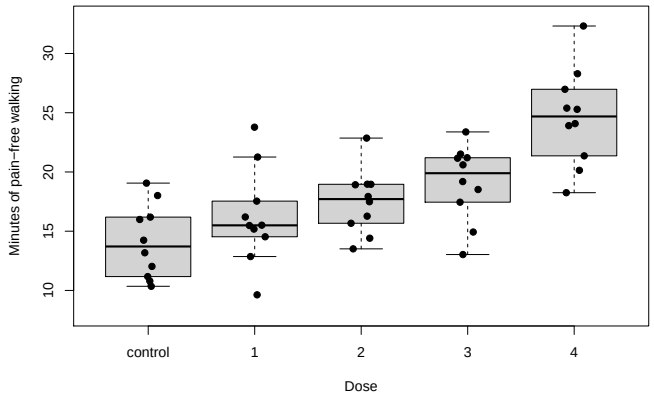

Figure 2.1: Boxplot for the Angina pectoris dose response data. Black points represent jittered data points giving artificial variation on the x-axis.

2.2 Micronucleus assay

In this data example the mutagenicity of hydrochinone is tested with a micronucleus assay based on male mice (Adler and Kliesch, 1990). An experimental design of four doses of hydrochinone, a negative (vehicle) control, and an active control, consisting of $25mg/kg$ cyclophosphamide, is used. Counts of micronuclei in polychromatic erythrocytes after 24h are taken as a measure for the potency to induce chromosome damage. The design is unbalanced with seven and four replications in the negative and positive control, and five replications for each of the dose steps.

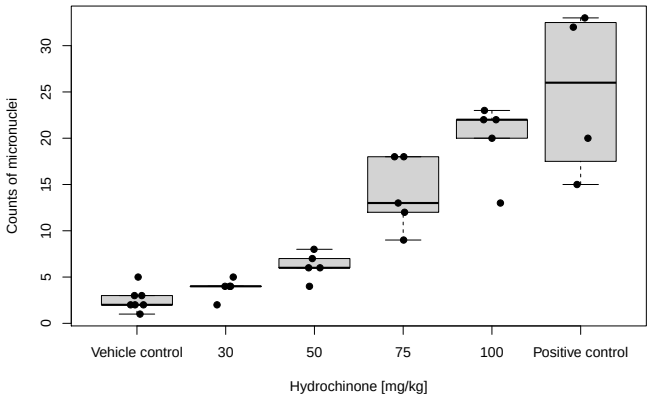

Figure 2.2: Boxplots for micronucleus assay data. Black points represent jittered data points giving artificial variation on the x-axis.

2.3 Cell transformation assay

This example data is provided by Thomas (2009), testing a chemical carcinogen by a cell transformation assay. Mouse embryo cultures (BALBc 3T3 cells) are treated with seven increasing doses of a carcinogen plus a solvent control. The original design consists additionally of an empty and a positive control, which are omitted for simplification. The cells treated with a carcinogen will not stop proliferation; therefore the number of type 3 foci, or cell clusters, are counted as an indicator for carcinogenicity. Ten dishes are used as replicates for each treatment level. The data is visualized in Figure 2.3a.

For only two dishes the number of foci in the solvent control is higher than zero. For a more extreme situation with zero foci counts in the control, a second data example is

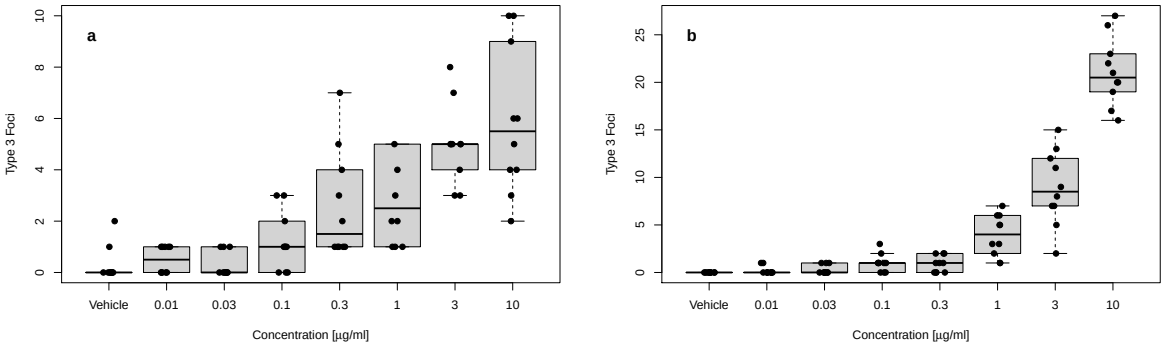

Figure 2.3: Boxplots for the data from two cell transformation assays (a, b). The dots at each treatment level visualize the actual observations; artificial variation is introduced on the x-axis to account for the discreteness of the data.

provided with the same design, testing a different carcinogen, shown in Figure 2.3b.

2.4 Liatrozole dose response study

A dose range case-control study was conducted to determine the lowest effective oral dose of liatrozole in the treatment of psoriasis vulgaris (Berth-Jones et al., 2000). Three doses of liatrozole and a placebo are randomly assigned to 116 patients, counting the number of individuals, where a marked improvement was observed. All dose groups are assigned with nearly the same number of patients, from 33 to 36 individuals. The resulting $2 \times k$ table is shown as mosaicplot in Figure 2.4.

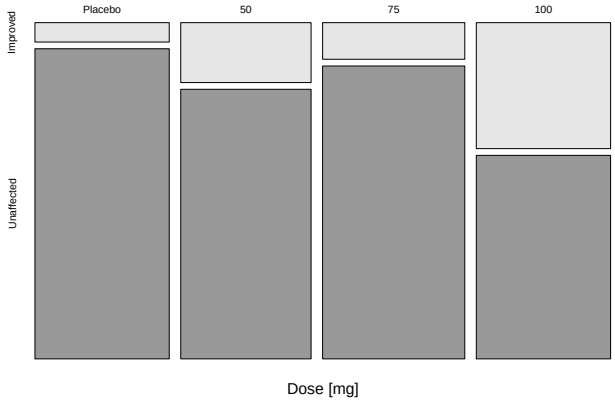

Figure 2.4: Mosaicplot for the liatrozole dose range data.

2.5 Fetal deaths after exposure to boric acid

This dataset represents results from a US National Toxicology Program (NTP, 1989) study of developmental toxicity in CD-1 Swiss mice, taken from Bailer and Piegorsch (1997). After female mice were exposed to 3 doses of boric acid in feed (0.1%, 0.2%, and 0.4%) and a control, the number of dead fetuses and litter sizes were recorded for each animal. The data is shown in Figure 2.5.

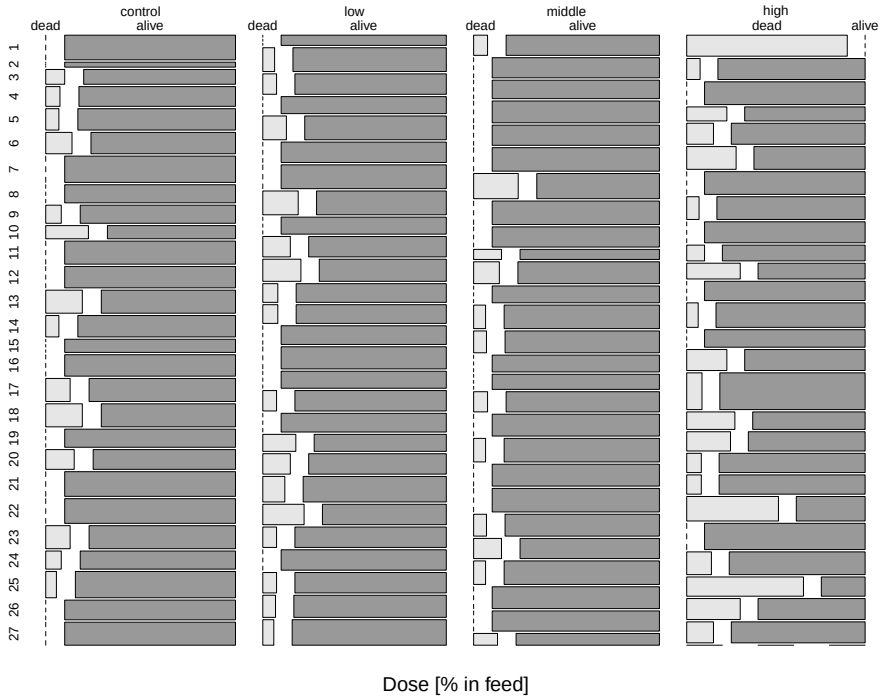

Figure 2.5: Mosaicplot for the fetal deaths data. The dotted lines indicate that no event was counted in this category.

Chapter 3

Methods

3.1 Parametric models

As a simple example, a vector of observed values $\boldsymbol{y} = (y_1, \ldots, y_n)$ can be assumed, which can be considered as a value taken on by a random variable Y. The goal is to make use of $\boldsymbol{y}$ in drawing conclusions on the unknown distribution $F(\cdot)$ of Y. A set of all possible choices of F, denoted as $\mathcal{F}$, is called a statistical model. In the special case, when all elements of $\mathcal{F}$ are of the same mathematical form, identified only by the different specification of parameters θ, where for any fixed θ, $F(\cdot\,;\theta)$ is a distribution function whose support is a subset of $\mathbb{R}^n$. The statistical model $\mathcal{F}$ can also be defined as a set of density functions

$$\mathcal{F} = \left\{ f(\cdot\,;\theta) : \theta \in \Theta \subseteq \mathbb{R}^k \right\} \tag{3.1}$$

for some density function f, and k being a positive integer.

3.1.1 Likelihood

Assuming a parametric model $\mathcal{F}$ parameterized by a fixed and unknown θ, the likelihood is the probability of the observed data considered as a function of θ;

$$L(\theta) = L(\theta; y) = c(y) f(y; \theta), \tag{3.2}$$

where $c(y)$ is a positive constant independent of θ.

Therefore the likelihood includes all information about a postulated statistical model. Information about the population of interest is supplied by a sample of observations, which is assumed to be representative for the whole population. Hence, the likelihood is calculated conditional on a random data sample, the only unknown variables are the coefficients θ of the model. By finding the θ, which maximizes the likelihood, the residuals of the models are minimized, and a parameter estimate $\hat{\theta}$ is found, being an unbiased estimate based on the assumption of the model. (Reid, 2000)

3.1.2 Generalized Linear Models (GLM)

A simple one way layout with a vector of observations $\boldsymbol{y} = (y_i)$ with $i = 1, \ldots, N$ as a realization of a random variable $\boldsymbol{Y}$ is assumed. The design of an experiment can be specified by a matrix of covariates $\boldsymbol{X} = (x_{ij})$, with $j = 1, \ldots, k$. Corresponding to this design matrix, an unknown parameter vector $\boldsymbol{\beta} = (\beta_j)$ contains the model coefficients of interest. These parameters can be estimated by minimizing the residuals $\boldsymbol{\epsilon} = (\epsilon_i)$ for a given deviance function. The model can then be written as

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{3.3}$$

The linear predictor $\boldsymbol{\eta}$ is calculated on a given link function $\tau(\cdot)$, for which the inverse can be used to obtain the predictions on the original scale

$$\boldsymbol{\eta} = \tau(\boldsymbol{\mu}). \tag{3.4}$$

Under the assumption that each component of $\boldsymbol{Y}$ has a distribution in the exponential family, the density takes the form

$$f_Y(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \tag{3.5}$$

for specific functions $a(\phi), b(\theta), c(y, \phi)$. As the parameters should be estimated conditional on an observation y the log likelihood function can be written as

$$l(\theta, \phi; y) = \log\left(f_Y(y; \theta, \phi)\right). \tag{3.6}$$

The parameter vector $\hat{\boldsymbol{\beta}}$ can be estimated by maximizing the (log) likelihood $l(\boldsymbol{\mu}; \boldsymbol{y})$ or equivalently minimizing the scaled deviance

$$D(\boldsymbol{\mu}; \boldsymbol{y}) = 2\left(l(\boldsymbol{y}; \boldsymbol{y}) - l(\boldsymbol{\mu}; \boldsymbol{y})\right) \tag{3.7}$$

with respect to $\boldsymbol{\mu}$. For a more detailed description see for example McCullagh and Nelder (1989), or Nelder and Wedderburn (1972).

3.1.3 Parameterization

There are several ways available to specify the design matrix $\boldsymbol{X}$ with the only requirement of being full of rank. Numerical covariates can be plugged into the design matrix, e.g. estimating the slope in a linear regression, where the intercept can be represented by a vector of ones. If categorical variables are of interest, they may be represented by a dummy coding of zeros and ones, corresponding to the estimation of means and/or differences between means of different factor levels. For example in an experiment with one factor having three factor levels, three treatment means may be estimated, but also the parameterization by one treatment mean and the difference to this mean for the two other factor levels is possible. In both cases the same predictions are obtained.

3.2 Deviance profiles

If a model consists of only a single parameter, e.g. estimating the mean of a sample of normal distributed observations, the deviance can be easily calculated conditional on this fixed parameter β from $\hat{\beta}$ by

$$d(\beta) = D\left(x_i\beta, y_i\right). \tag{3.8}$$

If, according to Chen and Jennrich (1996), interest lies in a real valued function $g(\beta)$, for any p in range of g, let $\hat{\beta}_p$ be the argument that maximizes $L(\beta)$ given $g(\beta) = p$. Viewed as a function of p, $d(\hat{\beta}_p)$ is called a deviance profile. In this case, g is only a function involving a single parameter.

For models with multiple parameters it is common to investigate the change in deviance separately for each single component of a parameter vector; thus, a reduction in dimension is needed, which can be defined by an adequate choice of g. The decrease of complexity for a function g is achieved, e.g. by calculating the deviance conditional on one parameter β_j, treating all others $\beta_{j'}$ as nuisance parameters. Hence, for a parameter vector $\boldsymbol{\beta} = (\beta_j)$,

the deviance profile is found by $d(\hat{\boldsymbol{\beta}}_p)$, with $\hat{\boldsymbol{\beta}}_p$ being the argument that maximizes $L(\boldsymbol{\beta})$ given $g(\boldsymbol{\beta}) = p$.

If the nuisance parameters are uncorrelated to the parameter of interest, their conditional estimates will be simply found at the unconditioned MLEs. Otherwise, $d(\hat{\boldsymbol{\beta}}_p)$ corresponds to the minimum deviance for the specific function g.

Instead of investigating the deviance statistic directly, Bates and Watts (1988) suggest to use the transformation

$$r(\hat{\boldsymbol{\beta}}_p) = sign(p - \hat{g})\sqrt{d(\hat{\boldsymbol{\beta}}_p)} \tag{3.9}$$

to the signed root deviance profile (SRDP), or signed log-likelihood ratio statistic (Barndorff-Nielsen, 1986). This statistic includes the additional information about the direction of the difference between each component of $\boldsymbol{\beta}_p$ and $g(\hat{\boldsymbol{\beta}})$. While Bates and Watts only construct their profile-t-plots for coefficients of non-linear regression models assuming a Gaussian distributed response, Chen and Jennrich (1996) generalized this idea to arbitrary functions of parameters in general maximum likelihood analysis.

Plotting the function $r(\hat{\boldsymbol{\beta}}_p)$ can reveal a large amount of information about the parameters in a model. Even for complex problems these profile plots represent an accurate summary for the observed data, illustrated directly for a parameter of interest. As the profile likelihood can be interpreted also as a likelihood, any information about the model and the sample of observations, that is, the uncertainty and distribution of a parameter is visualized. Chen and Jennrich (2002) even propose to derive transformations of the data to provide simple functions, which produce a nearly linear SRDP.

3.2.1 Observed information

The likelihood has a direct relationship to the available information from the model. If the likelihood is characterized by narrow curves, the parameters can be estimated more exactly with more information about the location of the MLE; at a larger width of the likelihood curve, less information is available. As a measure about the curvature of the log-likelihood the observed Fisher information function $j(\beta) = -\partial^2 l(\beta)/\partial\beta^2$ can be used. By plugging the maximum likelihood estimates into this observed information function, the observed Fisher information $j(\hat{\beta})$ is obtained. Instead of using the information about

the observations actually made, the expected Fisher information $i(\hat{\beta})$, can be used to give a generalization about the whole population (Lindsey, 1996), representing the $k \times k$ variance-covariance matrix for $\boldsymbol{\beta}$.

As it is shown in the Appendix (A.1) the Gaussian deviance is calculated by summing up the squared residuals. Hence, the deviance is increasing equally around the MLE as a quadratic function. The curvature of the deviance function can be completely covered by this quadratic function, and therefore summarizing the data by mean and standard errors leads to exact inference. In terms of a profile, the quadratic approximation (QA) corresponds to the well known Wald statistic

$$t(\beta) = j(\hat{\beta})^{\frac{1}{2}} \left(\hat{\beta} - \beta \right), \tag{3.10}$$

described for example in Lindsey (1996). Instead of using the expected Fisher information as in the original version by Wald (1949), the observed Fisher information is used here. The square of the QA profile has it's minimum at the MLE, increasing quadratically to both sides; hence, $t(\beta)$ is a linear function of β.

3.2.2 Linear contrasts of GLM parameters

In some situations the parameters of a model are not directly of interest, but linear combinations of them. These linear combinations of parameters can be presented as contrast matrices $\boldsymbol{C} = (c_{mj})$, $m = 1, \ldots, M$, multiplied with the parameter vector β_j

$$\psi_m = \sum_{j=1}^{k} c_{mj} \beta_j. \tag{3.11}$$

Some examples of choosing useful contrasts are presented in Hochberg and Tamhane (1987) or Bretz et al. (2001). There are several popular contrasts available, which are specifically discussed in the literature; the most famous ones considering comparisons to a control (Dunnett, 1955) and the all-pairs comparisons (Tukey, 1983). Additionally, trend tests are available based on Williams-type contrasts (Bretz, 2006). This contrast considers several convex alternatives, comparing the first with the last group, the first with the pooled two last, and so on until the first group is compared to a pooled estimate for all following groups. Detection of a changepoint by multiple contrasts (Hirotsu and Marumo, 2002) can be performed by dividing the parameters into all combinations of

two interconnected groups. Exemplary construction of these contrasts applied on five parameters is shown in the Appendix (A.2).

Also for the new parameter vector $\boldsymbol{\psi}_p = (\psi_m)_p$ deviance profiles can be constructed by calculating $r(\hat{\boldsymbol{\psi}}_p)$. The transformation of the parameter vector by the contrast matrix can be simply interpreted as a set of specific functions g_m of the model coefficients. But here, the additional problem occurs, when estimating a deviance statistic conditional on a parameter ψ_m, multiple solutions are existing, as the parameter linear combinations can be constructed by multiple sets of model coefficients. Therefore, a constraint can be introduced in the optimization algorithm, finding $k-1$ model parameters that minimize the deviance, conditional on, for example, a redundant parameter β_1, by

$$\psi_m - \sum_{j'=2}^{k} c_{mj'}\beta_{j'} = c_{m1}\beta_1, \tag{3.12}$$

where $\beta_{j'}$ are treated as nuisance parameters. It is of direct importance, that the coefficient for the redundant parameter c_{m1} is not equal to zero, as then, multiple solutions are available.

3.2.3 Profiling for the angina data

The angina dataset introduced in Section 2.1 consists of a continuous response measured for five dose groups. When assuming the duration of pain-free walking following a Gaussian distribution, for each group a mean parameter can be estimated, minimizing the squared residuals. If comparisons of all four dose groups versus the control are of interest, the contrast matrix

$$c_{mj} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

can be used to transform the estimated parameter vector. For the resulting parameter estimates of interest, that is, the difference between averages of the doses compared to the control mean, the deviance profiles are calculated. By minimizing the deviance conditional on each linear combination of means, single profile curves for each new parameter are obtained, which is demonstrated on the difference of the first dose average to the control

in Figure 3.1. All parameters with contrast coefficients of zero are fixed at their MLEs. The remaining three contrasts for the comparison of each dose group to the control result in similar profiles, only showing a shift in the location of the MLEs.

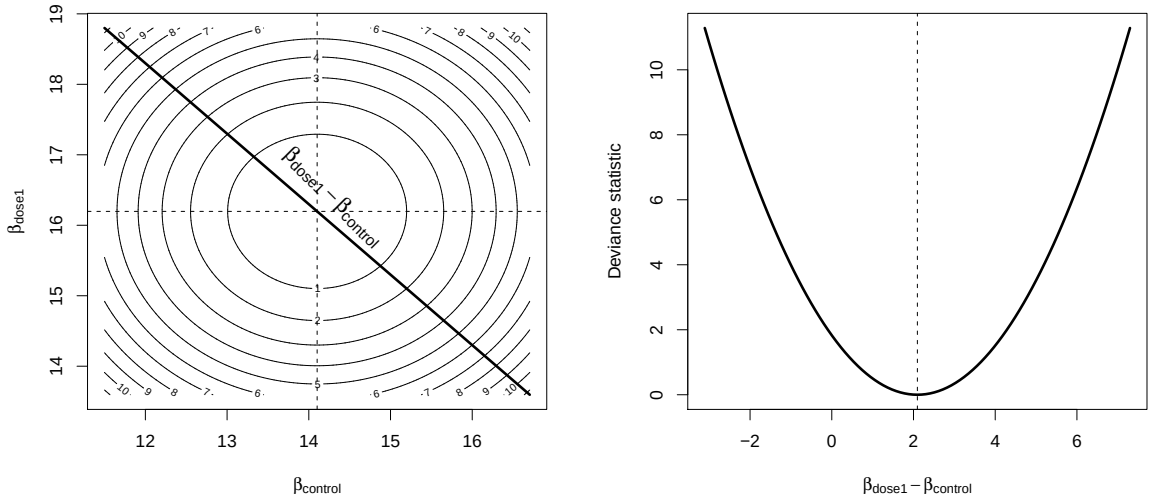

Figure 3.1: *left*: Two-dimensional deviance profile for the parameters coding for the means of dose 1 and the control in a linear Gaussian model for the angina dataset. The bold intersecting line represents the deviance profile for the difference of both parameters. *right*: The profile deviance curve for the difference of means of dose 1 and the control, presented as bold line in the left contour plot.

As a Gaussian linear model with the assumption of homogeneous variances is used, and the parameters for Dose 0 and Dose 1 are uncorrelated, the deviance increases symmetrically with increasing distance to the coordinate of the two MLEs. If the difference between these two parameters is increased by some constant d, the minimum deviance conditional on this difference is reached for $\left(\hat{\beta}_2 + d/2\right) - \left(\hat{\beta}_1 - d/2\right) = \hat{\psi}_1 + d$, giving equal weight to each of the original parameters; hence, the profile for the difference of the two means is found by a bisecting line in the two-dimensional deviance contour. This profile is a quadratic function, explaining the deviance statistic by the parameter for the difference of means.

3.2.4 Profiling the cell transformation assay data

The response in the cell transformation assay data (Section 2.3) are counts of cell clusters. These counts may be modeled adequately by assuming them to be Poisson distributed,

estimating mean parameters for each dose group. As multiple dishes per dose group are evaluated, being an additional source of variation, the variance might be underestimated assuming a simple Poisson GLM; therefore a quasi-Poisson model might be assumed, described more detailed in Section 3.8. This model implies a loose variance-mean relationship, scaling the variances by a single parameter estimated from the data. Likewise to the profile for the angina data, the profiling is shown for the difference of the lowest dose ($0.01\ \mu g/ml$) to the control in Figure 3.2. The low number of counts in both groups cause a deviation from a quadratic profile; for all further comparisons of higher dose groups to the control this effect decreases as the number of counted events increases.

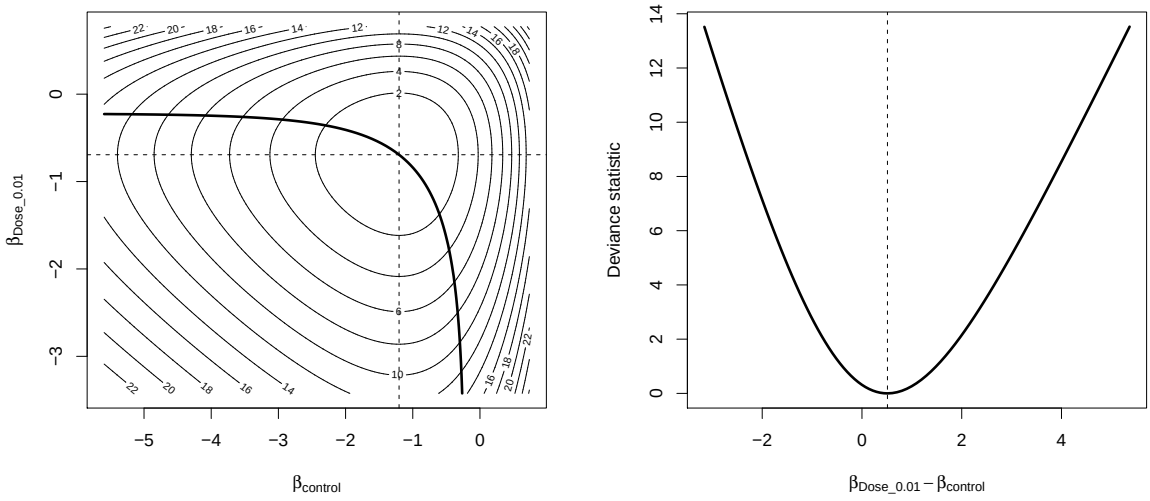

Figure 3.2: *left*: Two-dimensional deviance profile for the parameters coding for the means of dose 0.01 and the solvent control on a logarithmic scale in a quasipoisson GLM for the first cell transformation assay dataset. The bold intersecting line represents the deviance profile for the difference of both parameters. *right*: The profile deviance curve for the difference of means of dose 0.01 and the solvent control on the log link, presented as bold line in the left contour plot.

Due to the assumption of the variance dependence of the mean the deviance contour is not symmetric as in the Gaussian case. It follows, that the profile is not found easily on the bisecting line in the contour; a change in the difference between means arises not from an equal change of the primary parameters. Also the shape of the deviance profile function deviates from a quadratic shape, as the underlying Poisson distribution is not equal-tailed.

3.3 Hypothesis testing based on deviance profiles

The deviance statistic can be directly interpreted as likelihood ratio test

$$\Delta(p) = \frac{L(\hat{\beta}_p)}{L(\hat{\beta})}, \tag{3.13}$$

testing the single null-hypothesis $H_0 : \ p = g(\beta)$, as described in Chen and Jennrich (1996). Under the assumption, that $-2 \log \Delta(p)$, or equivalently $d(\beta_p)$ following approximately a χ^2 distribution with 1 degree of freedom, a test decision can be obtained. For Gaussian distributed data, an F distribution with adequate degrees of freedom can be assumed for the deviance test statistic.

If otherwise the signed root deviance statistic $r(\beta_p)$ is used to test the same hypothesis $H_0 : \ p = g(\beta)$, a standard normal distribution can be assumed; in case of Gaussian distributed data, a t-distribution with residual degrees of freedom might be a better choice at small samples (Bates and Watts, 1988).

In practice, the value of p can be chosen to define the limits of a rejection region of the test. The deviance statistic at $d(\hat{\beta}_p)$ is compared with a $1 - \alpha$ quantile of a χ^2- or F-distribution, with α defining the type-I-error rate for a single hypothesis. For the SRDP, $r(\hat{\beta}_p)$ is compared with a $1 - \alpha$ quantile of the standard Normal- or t-distribution.

3.3.1 A shortcut

As discussed in Section 3.2.1 the deviance profile can be approximated by a quadratic function. By calculating the inverse of the information matrix, estimates of standard errors of the parameters are obtained, representing the curvature of the quadratic approximation to the likelihood. Hence, the deviance profile is summarized by it's minimum and the curvature of a quadratic function, and also a test can be based on both of these measures. This leads to the well-known z-statistic

$$\hat{T} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \tag{3.14}$$

with $\hat{\sigma}_{\hat{\beta}}$ being the estimated standard error of $\hat{\beta}$. As for Gaussian distributed data the deviance profile and it's quadratic approximation are similar, the same test results are obtained for both profiles.

3.3.2 Multiple hypotheses testing

Instead of performing only one test, based on a single profile, several functions of parameters g_m for $m = 1, \ldots, M$ might be of interest. Hence, the corresponding set of null-hypotheses can be stated as

$$H_0 : \quad \bigcap_{m=1}^{M} p_m = g_m(\boldsymbol{\beta}). \tag{3.15}$$

If separate univariate distributions are assumed for each test statistic, only a local type-I-error rate can be controlled. If the control of a global error rate, e.g. controlling the family-wise-error rate, is of interest, a multivariate distribution has to be assumed for the vector of test statistics. Performing the relatively simple adjustment by Bonferroni, dividing α by the number of comparisons, assumes that all test statistics are uncorrelated. This is a quite strong assumption, as the set of g_m are all functions of the same parameter vector $\boldsymbol{\beta}$; additionally, only if the distribution of the parameters implies the independence of mean and variance, this assumption is reasonable.

As an alternative, a multivariate distribution with the known correlation structure of the test statistics can be assumed. Obviously, there is the problem of obtaining such a correlation structure. One simple solution is to estimate this correlation directly from the observed data. This approach implies, that the underlying distribution of the test statistics depend on the sample of observations; therefore, it is assumed, that accurate estimates of the correlation structure are available. Hothorn et al. (2008) state, that this approximation of the underlying distribution by a normal distribution with an estimated shape can be made even for relatively small sample sizes. For Gaussian distributed data in a linear model, assuming homogeneous variances, this assumptions lead to exact inference. With $\hat{\beta}_j$ being an estimate of β_j, $\hat{\boldsymbol{\Sigma}}_j$ being an estimate of $cov(\hat{\beta}_j)$ and assuming that a multivariate central limit theorem holds, then

$$\hat{\beta}_j \overset{a}{\sim} N(\beta_j, \hat{\boldsymbol{\Sigma}}_j). \tag{3.16}$$

The next problem is the estimation of the correlation structure from the data. It might be quite difficult to obtain any adequate information about the desired correlation directly from a complete deviance profile; but the simplification of the problem by a quadratic approximation is well-known and directly available. Therefore, as described in Hothorn

et al. (2008) the estimated variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_j = (\hat{\sigma}^2_{j,j})$ can be standardized to an estimated correlation matrix. Quantiles and adjusted p-values from an estimated normal distribution can be calculated according to Bretz et al. (2001).

For all comparisons a single quantile is calculated, allocating the available global type-I-error α equally on all simultaneously tested hypotheses. This method is focussed on the maximum of test statistics to control the family-wise error rate in the strong sense. This allows additional conclusions about the rejection of any local hypothesis at the global type-I-error level.

Due to the numerical availability of quantiles and percentiles only for a multivariate normal (and t-) distribution and not for a multivariate χ^2 distribution, the use of the signed root deviance profile is proposed. Similar test results may be produced for deviance profiles generating χ^2 quantiles by taking the square of the quantiles calculated from a multivariate normal distribution.

3.3.3 Multiple contrast tests

When performing tests for linear contrasts of model parameters, the profiles for these linear combinations of parameters can be used directly to construct the test statistics. To estimate the shape of the underlying distribution, also the quadratic approximation to the profile can be used. Additionally to the correlation between parameters of the full model, the correlation due to the specified contrasts have to be taken into account. Practically this means, that the variance-covariance matrix estimated from the model can be multiplied with the predefined contrast matrix

$$\hat{\boldsymbol{\Sigma}}_{\hat{\psi}} = \boldsymbol{C}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}' \tag{3.17}$$

to obtain adequate variance estimates for the contrast parameters. By investigating comparisons of several parameters specified by a contrast matrix, the resulting parameters of interest are most likely correlated. Considering the correlation due to the contrasts we can assume

$$\boldsymbol{C}\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{C}\boldsymbol{\beta}, \hat{\boldsymbol{\Sigma}}_{\hat{\psi}}), \tag{3.18}$$

see Hothorn et al. (2008).

To perform the multiple test, a vector of test margins p_m is specified; the corresponding signed root deviance statistics $r((\hat{\psi}_m)_{p_m})$ are compared with a $1-\alpha$ quantile of a multivariate normal or t-distribution with the estimated correlation structure.

When using the shortcut over the quadratic approximation to construct associated Wald statistics, these are constructed by

$$\hat{T}_\psi = \frac{\hat{\psi}_m}{\hat{\sigma}_{m,m}^{(\hat{\psi})}}. \tag{3.19}$$

For a linear Gaussian model with homogeneous variances the correlation of test statistics exclusively depends on the specified contrast. Accordingly, an exact test is constructed for this special case under assumption of Equation 3.18.

3.4 Confidence intervals based on profile deviance

When believing in a true but unknown model parameter, estimating it by minimizing the deviance should provide the best idea about the location of the parameter, based on a representative random sample of observations. But this estimator is a random variable, showing some variability for different sets of observations; therefore the standard error can be used to obtain some uncertainty measure about the reliability of the MLE. Standard errors indicate the uncertainty about the parameter estimate, but are not directly connected to the underlying true parameter in the model. To obtain conclusions about the true parameter, some assumption about the distribution of the parameter estimate has to be made. By calculating confidence intervals, given the distribution of the parameter estimate, some probabilistic information can be inferred about the location of the true parameter. Here, only frequentist inference is discussed, in a sense, that the true parameter is located within the estimated confidence interval limits in $100(1-\alpha)\%$ of replicated samples from the same population.

The calculation of these limits is based on a random sample of observations; therefore the limits of the confidence intervals are also random variables. Hence, the confidence intervals give only a frequentist view on the location of the true parameter, that is for $N(1-\alpha)$ out of N random samples the parameter is located within the particular confidence limits.

3.4.1 Univariate intervals

Confidence intervals based on a deviance profile can be constructed by inversion of the likelihood ratio test (Chen and Jennrich, 1996). Under assumption of a χ^2 distribution for the deviance statistics, the confidence limits are found at the parameter values at which the test is just about to reject the null-hypothesis. That is done by searching the profile for a statistic that equals a corresponding quantile of the underlying distribution. For signed root deviance profiles an approximate normal distribution can be applied, whereas for Gaussian model the assumption of F- and t-distributions are suitable. The confidence interval based on the deviance profile is defined by

$$CI_{deviance} = \left\{p : d(p) \leq q_{1-\alpha/2}\right\} \tag{3.20}$$

where $q_{1-\alpha/2}$ is the $1-\alpha$ quantile of a χ^2 distribution with 1 degree of freedom. A corresponding interval based on the signed root deviance profile is obtained by

$$CI_{SRDP} = \left\{p : z_{\alpha/2} \leq r(p) \leq z_{1-\alpha/2}\right\} \tag{3.21}$$

with z being a quantile of the standard normal distribution.

Confidence intervals can also be constructed based on a quadratic approximation to the likelihood, see for example Lindsey (1996). The shortcut of calculating standard errors as a measure of profile curvature provides confidence limits

$$\left[\hat{\delta}_l, \hat{\delta}_u\right] = \left[\hat{\beta} \pm z_{1-\alpha/2}\hat{\sigma}_{j,j}\right] \tag{3.22}$$

with the interval

$$CI_{QA} = \left\{p : z_{\alpha/2} \leq t(p) \leq z_{1-\alpha/2}\right\}. \tag{3.23}$$

The construction of univariate confidence intervals is illustrated in Figure 3.3 based on the profile for the difference between the first dose versus the control of the angina dataset. As a Gaussian GLM is assumed, the deviance profile is a quadratic function, thus the shortcut by adding and subtracting a quantile times the estimated standard deviation is available.

An important feature of deviance profiles is their transformation invariance (Lindsey, 1996). If a confidence interval for a parameter on a transformed scale is needed, it can be obtained by transforming the confidence limits for the original parameter without loss of

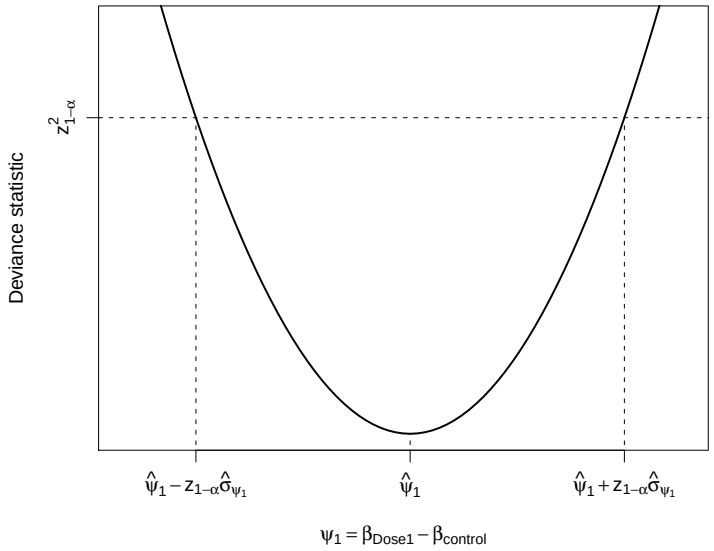

Figure 3.3: Construction of the univariate confidence intervals for the difference of means of doses 1 and 0 for the angina dataset assuming a Gaussian linear model. $\hat{\psi}_1$ represents the parameter estimate for the difference. By searching for the intersections of the cutoff value $z^2_{1-\alpha}$ with the deviance profile, estimates of upper and lower confidence limits can be found. $z^2_{1-\alpha}$ represents the squared $1-\alpha$ quantile of a standard normal distribution. The shortcut $\hat{\psi}_1 \pm z_{1-\alpha}\hat{\sigma}_\psi$ is only available for the quadratic approximation to the deviance.

information. As the curvature of the quadratic approximation, inferred from the observed information function, depends on the scale of the corresponding parameter, transformation invariance is not given for these confidence intervals (Section 3.5.1).

3.4.2 Confidence regions

Instead of single parameter models, most models of interest are containing multiple parameters; therefore a restriction to single profiles without considering the correlation between the parameters is not recommended. As for a single parameter a deviance curve can be used to obtain all information about a parameter, for multiple parameters a deviance surface in a k-dimensional space has to be considered. Treating the vector of parameter estimates $\hat{\beta}_j$ as point in this k-dimensional space, a confidence region can be defined around it, which includes the true parameter vector with a predefined error rate. This confidence region can be calculated by choosing an adequate cut-off value, similarly to the univariate deviance intervals. Here, also a quantile of the χ^2 distribution can be used,

resulting in a convex space, which includes the true parameter coordinate in $N(1-\alpha)$ out of N random samples of the same population of observations. This confidence region also corresponds directly with an inversion of a likelihood ratio test.

Confidence regions for the first two components of the parameter vector of a Gaussian linear model and a quasipoisson GLM based on the angina and cell transformation assay example are presented in Figure 3.4. The parameters ψ_1 and ψ_2 are coding for the difference of the means of the first two dose groups with the control group. Both differences share the comparison to the same control group; therefore the two parameters are correlated $\rho = 0.5$, resulting in elliptical confidence regions.

Of course, a corresponding confidence region can also be constructed on the quadratic approximation to the deviance; but obviously a nice shortcut as for univariate confidence intervals is not available, as a complete parameter space is of interest and not only two single confidence limits.

If linear combinations of the model parameters are of interest, in a simple way the complete confidence region for the original model parameters can be transformed into the space of the new parameters of interest by multiplication with a predefined contrast matrix.

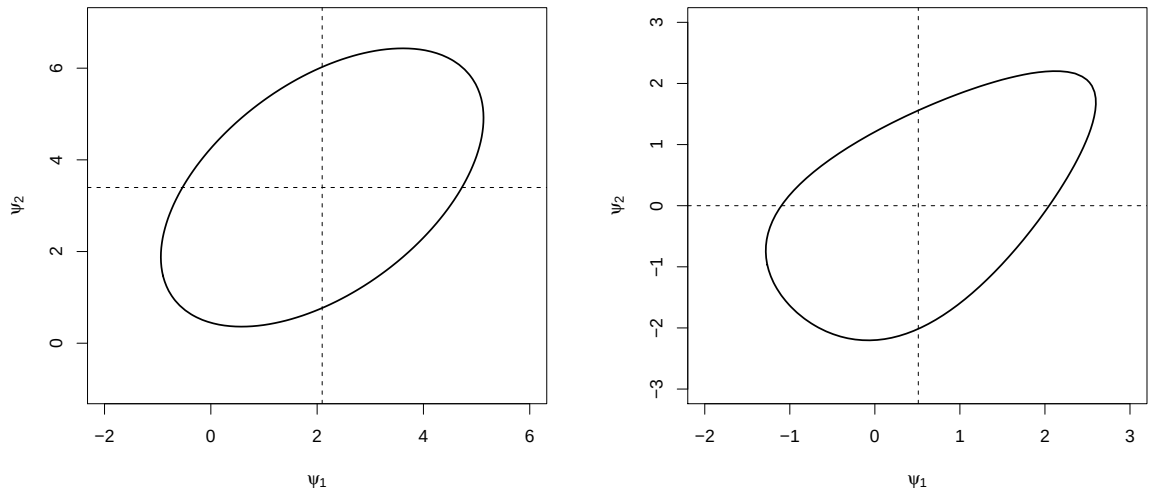

Figure 3.4: *left*: Confidence region for the differences between the means of the first two dose groups versus the control of the angina data example estimated in a Gaussian linear model. *right*: Confidence region for the differences between the means of the first two dose groups versus the solvent control, estimated on the log link in a quasipoisson GLM based on the cell transformation assay example.

3.4.3 Confidence intervals for multiple parameters

Interpreting parameters with respect to all other parameters in a model, based on a confidence region, is quite difficult, as for complex problems, looking at features in a k-dimensional parameter space, is quite impossible to grasp. It is easier to look at single confidence intervals separately for each parameter in the model. The problem is here to summarize the information of the complete confidence region to single confidence intervals on the scale of each single parameter. One possibility to obtain confidence intervals for each parameter separately, is to project the minimum and maximum values of a confidence region on the scale of a particular parameter. This procedure matches the inversion of single hypotheses tests each tested at a local type-I-error α.

Instead of calculating minimum and maximum values for a complete confidence region, a deviance profile conditional on a single parameter, treating all others as nuisance parameters, as described in Section 3.2 includes all this information needed. By minimizing the deviance for the nuisance parameters conditional on one parameter of interest, the resulting profile curve shows directly the projections of the minimum and maximum limits of a corresponding confidence region on the scale of interest. Therefore, separate profile curves for each parameter also reflect the correlations between parameters since they are treated as nuisance parameters and optimized simultaneously. This also holds for profiles for multiple linear combinations of parameters due to the transformation invariance of the profiling. Confidence intervals for multiple parameters are in the same manner constructed as the univariate intervals by choosing an adequate cut-off value for the deviance statistic. The projections of the minimum and maximum on each single axis of the coordinate system yields then a rectangular confidence region.

If a parameter vector is interpreted as coordinate vector in a multidimensional parameter space, the cut-off value for a complete confidence region can be chosen as quantile of a univariate χ^2 distribution. As described in Section 3.4.2 this confidence region will contain the true coordinate vector with probability $1-\alpha$. Therefore, the corresponding confidence intervals, obtained from the projections on the coordinate axes, will also contain each particular true parameter with probability $1-\alpha$.

These multiple univariate confidence intervals are calculated, e.g. for the signed root

deviance profile, by

$$CI_{SRDP} = \left\{ p_m : z_{\alpha/2} \leq r(p_m) \leq z_{1-\alpha/2} \right\}, \tag{3.24}$$

resulting in a rectangular confidence set.

Instead of calculating single confidence intervals each at a local α level, an inversion of the multiple test, discussed in Section 3.3.2, can be applied. The resulting confidence intervals will maintain a global error rate, that is, with a probability of $100(1-\alpha)\%$ the true parameter will be contained in each of the intervals. The simultaneous confidence intervals are calculated in a similar way as the multiple univariate ones, but with an adequate critical value likewise to the corresponding multiple test. Choosing the same quantile for all confidence intervals allocates an equal amount of the global α to each of the single, local confidence intervals. By the projection of the minimum and maximum values in the confidence region onto the axes of the coordinate system, a rectangular confidence set is obtained, as illustrated in Figure 3.5.

For uncorrelated parameters the simultaneous confidence level is adequately controlled by α adjustment according to Bonferroni. With an increased correlation of the parameters, the shape of the confidence region becomes more elliptical. A major source of this correlation is the choice of the contrast, applying multiple linear combinations on the same set of original model parameters. The corresponding profile curves summarize the confidence region by conditional estimation of the model parameters dependent on their weight in the linear combination.

3.5 Profile- vs. quadratic approximation based confidence intervals

For Gaussian distributed data, the deviance profile is a quadratic function and therefore similar to the quadratic 'approximation' to this profile. Hence, the confidence intervals based on the deviance profile will exactly match the ones of the quadratic function. For non-equal-tailed distributions, the quadratic approximation might not completely have the same curvature as its deviance profile. The difference between both profiles depends on the skewness of the distribution of the data and the sample size. With increasing

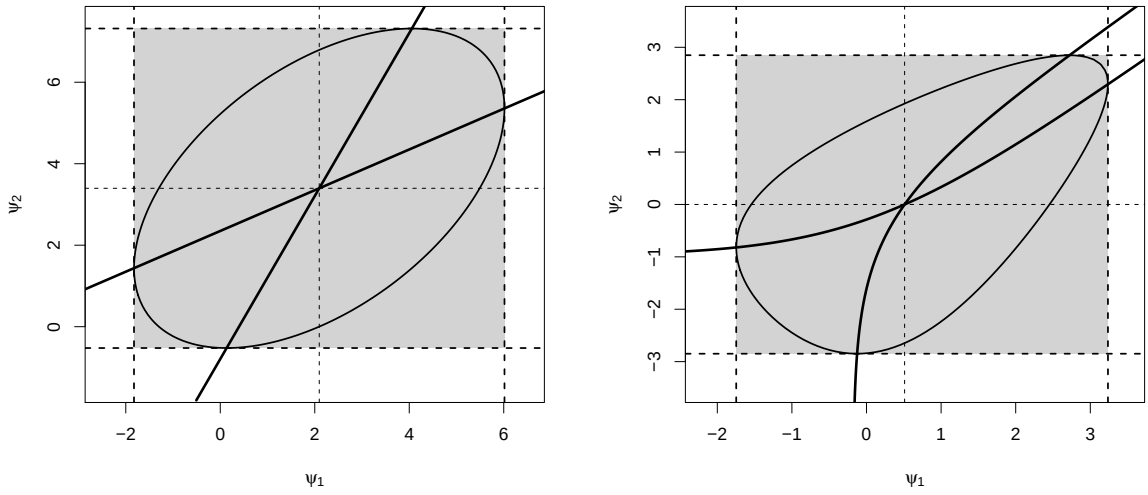

Figure 3.5: Confidence regions for two components of a parameter vector referring to a common intercept are represented by an elliptical contour. The two bold lines crossing the point estimate represent the location of the profile curves for each single parameter component ψ_1 and ψ_2. The gray area represents the confidence set for ψ_1 and ψ_2 spanned by the intersections of the profile curves with the limits of the confidence region. The dotted lines illustrate the projection of the confidence limits and the point estimate to the axes of the coordinate system. *left*: Confidence region and corresponding confidence set for the difference of means for dose 0 and 1 to the control estimated in a Gaussian linear model based on the angina data example. *right*: Confidence region and confidence set for the difference of means for dose 0.01 and 0.03 to the solvent control, estimated on a logarithmic scale in a quasipoisson GLM based on the cell transformation assay example.

sample sizes the differences between the deviance profile and the quadratic approximation decreases; for example the Poisson distribution can be adequately approximated by a normal distribution, if a large enough number of events were counted.

3.5.1 Transformation invariance

To improve the quadratic approximation to the likelihood the parameters can be estimated on a transformed scale. In the case of generalized linear models these are defined by the link function (Eq. 3.4). If, for example, count data are analyzed in a Poisson GLM, the variance is assumed to be completely determined by the mean. Expecting a large variance for a large number of counted events and a smaller variance at less events counted, estimating parameters on a logarithmic scale is an obvious choice. The effect of this transformation is illustrated on the micronucleus assay example (Section 2.2) in Figure

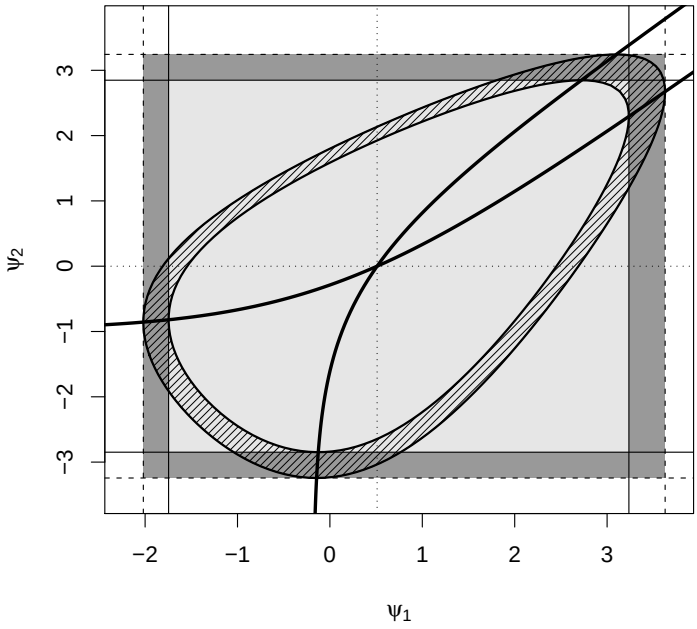

Figure 3.6: Construction of the univariate confidence intervals for the difference of means of doses 1 and 0 for the angina dataset assuming a Gaussian linear model. $\hat{\psi}_1$ represents the parameter estimate for the difference. By searching for the intersections of the cutoff value $z^2_{1-\alpha}$ with the deviance profile, estimates of upper and lower confidence limits can be found. $z^2_{1-\alpha}$ represents the squared $1-\alpha$ quantile of a standard normal distribution. The shortcut $\hat{\psi}_1 \pm z_{1-\alpha}\hat{\sigma}_\psi$ is only available for the quadratic approximation to the deviance.

3.7. Comparing both profiles a small improvement of the quadratic profile can be found for the log link, especially for the upper confidence limits.

One problem of the transformation approach is, that the parameters on the transformed scale might not be of direct interest. A transformation of a confidence set back to the original scale of interest is not possible in any case. If a confidence interval for the difference of two parameters on a logarithmic scale is calculated, a transformation by taking the exponent of the difference will lead to the ratio of those two parameters. If the complete deviance profile is used to construct confidence intervals, this can be done on any scale of interest; but there are computational limits, as reaching convergence for a constrained likelihood optimization is easier on a canonical link. A discussion and simulation study about choosing a non-canonical link can be found in Czado and Munk (2000). The problem gets even more complex when looking at linear combinations of parameters. If linear contrasts are used to compare weighted groups of parameters, the

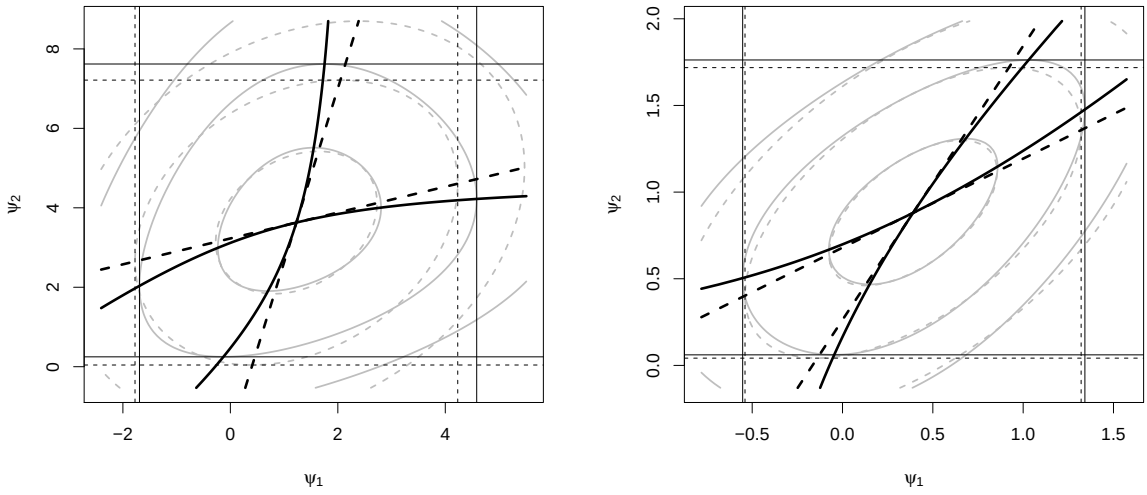

Figure 3.7: Two dimensional deviance contour plot for the differences of means for dose 0.01 and 0.03 to the solvent control estimated in a quasipoisson GLM based on the cell transformation assay example. The straight lines represent the deviance profile, whereas the dotted lines visualize its quadratic approximation. The lines crossing the point estimate represent the location of the profile curves for each single parameter component β_1 and β_2, with corresponding confidence set shown as thin black lines. *left*: Parameters are estimated on the identity link. *right*: Parameters are estimated on a logarithmic scale.

parameters are estimated on the specific link and pooled afterwards by multiplication with a contrast matrix. A second strategy is the definition of an adequate design matrix, estimating directly a parameter for the difference of pooled group means. In this case the pooling is done before transformation with a specific link, leading to a different model and therefore to a different parameter of interest.

A second problem is the dependence of the quadratic approximation on a distinct transformation. Estimating confidence limits for parameters on the original scale and on a transformed scale (using an adequate inverse of the transformation function) will lead to different results, as the curvature of the quadratic profile and thereby also the standard error estimates depend on the chosen transformation. Depending on the specified link function and a chosen linear contrast, the confidence limits are in some situations not fully supported by the likelihood.

3.5.2 Extreme Settings

With decreasing distance of the parameters to the boundary of the parameter space the difference between the deviance profiles and the quadratic approximation increases. Hence, if a parameter is located directly on the boundary of the parameter space, the difference between both approaches should be maximal. Again a simple example is presented by observing count data, being limited to observations of only positive integers. The second cell transformation assay dataset (Section 2.3) contains a control group, where for 10 replications no cell cluster was counted at all. Comparing the multiple dose groups to this control by confidence intervals for the difference of means estimated in a Poisson GLM is quite challenging. To enable an accurate estimation of parameters, a logarithmic transformation is applied, expanding the parameter space from $-\infty$ to $+\infty$. As the parameter estimates are found iteratively, a stopping rule can be introduced, searching for adequate estimates only within e.g. the interval $[-10; 10]$; thus, the control group estimate will be set to the value -10. The deviance profile and the quadratic approximation is illustrated for the comparisons of the first two doses versus the control in Figure 3.8.

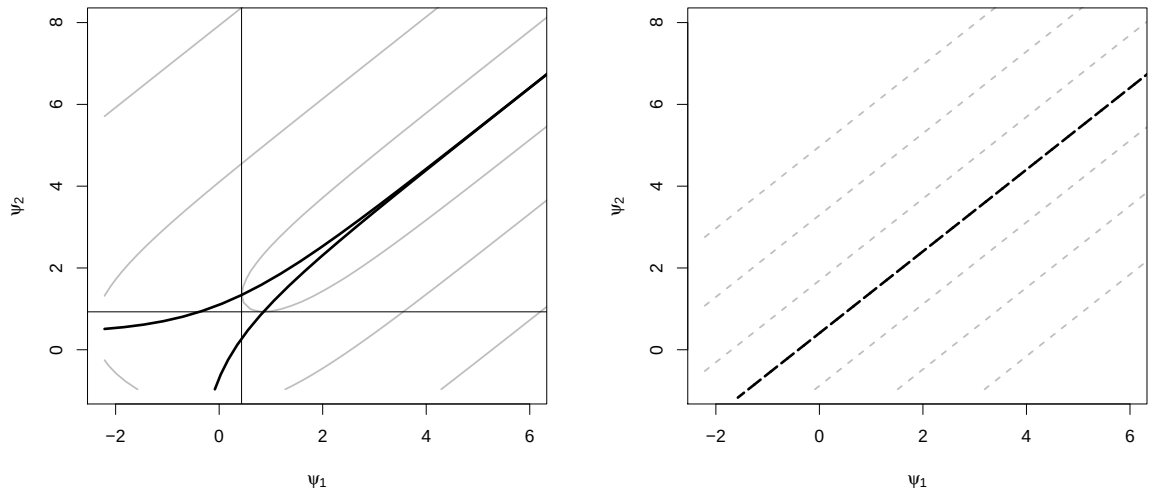

Figure 3.8: Two dimensional deviance contour plots for the differences of means for dose 0.01 and 0.03 to the solvent control estimated in a quasipoisson GLM based on the second cell transformation assay example. The bold lines represent the location of the profile curves for each single parameter component β_1 and β_2, with corresponding confidence set shown as thin black lines. *left*: Contour plot based on the profile deviance. *right*: Contour plot based on the quadratic approximation to the deviance.

Only lower interval limits are found for the profile deviance intervals, as the estimates for

the difference to the control are located at $+\infty$, or at 10 respectively. The quadratic approximation provides no confidence limit at all, as the standard error for these parameter estimates tend also to $+\infty$ and the quadratic function will be reduced to almost a straight line at a deviance of zero. This effect is known as the Hauck-Donner phenomenon, which was first described by Hauck and Donner (1977) for logistic regression.

It might be debatable, if the construction of confidence intervals with an approximate normal distribution as assumption for the parameters is adequate. At least, with the profile deviance approach a confidence interval is obtained, based only on the information of the dose groups, where some events are counted. The calculation of a confidence interval for a parameter, for which no useful information is available at all, that is for example if no event has been counted in any of the compared doses, is not applicable for deviance profiles.

Additionally, the construction of the profile in these situations is not computational stable. The parameter estimates of course depend on the artificial parameter space limitations and the algorithm used for constrained optimization.

3.6 Modifications of deviance statistics

By the calculation of the complete deviance profile instead of the computationally simple quadratic approximation, the resulting confidence intervals and tests should have more accurate properties at non equal-tailed distributed data. That is, the more skewed the distribution is, the more suitable the deviance profile becomes. On the contrary, the approximation of the statistic by a normal distribution changes for the worse; the more a profile deviates from a quadratic function, the more unrealistic an underlying normal distribution becomes. As the deviance profile is based on a first order approximation to the likelihood, modifications of the statistic by higher order approximations may improve their characteristics. Several methods are available dealing with higher order approximations; here only one of the most prominent methods, the Barndorff-Nielsen approximation based on saddle-point approximations (Barndorff-Nielsen, 1986), (Barndorff-Nielsen and Cox, 1979), is briefly discussed. This approach arises from an approximation to the density of

the maximum likelihood estimate

$$f(\hat{\beta}|a;\beta) \doteq c|j(\hat{\beta})|^{\frac{1}{2}} \exp\left(l(\beta) - l(\hat{\beta})\right) \equiv p^*(\hat{\beta}|a;\beta), \tag{3.25}$$

with c being a normalizing constant. The dimension of the sample space is reduced to a k-dimensional space of the parameter vector $\boldsymbol{\beta}$, by conditioning on a statistic a (Brazzale et al., 2007). a needs to be an ancillary statistic, that is, the distribution of a needs to be free of β. By the density of the maximum likelihood estimate, tail area approximations can be derived, leading to the modified signed root deviance statistic

$$r^*(\beta) = r(\beta) + \frac{1}{r(\beta)} \log\left(\frac{q(\beta)}{r(\beta)}\right). \tag{3.26}$$

For single parameter GLMs, $q(\beta)$ can be substituted by the Wald statistic $t(\beta)$ (Eq. 3.10). As $q(\beta)$ contains information about the curvature of the likelihood function at the MLE, the modified deviance statistic includes an adjustment for higher moments. Assuming a distribution of the exponential family in the GLM leads then to almost exact inferences, as described by Davison (1988). The detailed derivation of $r^*(\beta)$ is for example described in Brazzale et al. (2007, p.148). Confidence intervals and tests can be calculated in a similar way as for the deviance profiles using the modified statistic, with an exception near the MLE. Here, a singularity occurs, where both the Wald and the deviance statistic become zero. For confidence interval construction an acceptable solution to this problem is a spline interpolation at an appropriate interval around the MLE; a hypothesis test is not affected by this peculiarity.

3.6.1 Nuisance parameters

In multi-parameter models the reduction of the sample space to the k-dimensional space of the parameter becomes more complex, as additionally the effect of the nuisance parameters, for which the likelihood is maximized to obtain the profile for a fixed parameter, has to be considered. Not only the curvature of the likelihood at the MLE is needed, but also the change in curvature for the conditional MLE given the estimated nuisance parameters.

A parameter vector $\boldsymbol{\beta} = (\psi, \lambda)$ is introduced here, with parameter linear combination ψ for profiling and the nuisance parameters λ of secondary interest. The signed root

deviance statistic is calculated as described in Section 3.2, obtaining $r(\psi)$ by maximizing the likelihood for λ conditional on ψ. Additional to the Wald statistic with information about ψ an adjustment term for the nuisance parameters ρ has to be introduced by

$$q(\psi) = t(\psi)\rho(\psi, \hat{\psi}). \tag{3.27}$$

The Wald statistic is given by

$$t(\psi) = j_p(\psi)(\hat{\psi} - \psi) \tag{3.28}$$

with the observed Fisher information of the profile log-likelihood $j_p = -l_p''(\psi)$. The information about the nuisance parameters

$$\rho(\psi, \hat{\psi}) = \left(\frac{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|}{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|} \right)^{\frac{1}{2}} \tag{3.29}$$

contains the determinants of the observed Fisher information function for the full model $|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|$ and the constrained fit $|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|$. Hence, all components, which need to be calculated, are available by performing a constrained and the full parameter optimization. The singularity around the MLE is still a small problem, which can be solved by an adequate spline interpolation.

3.6.2 Multi-parameter profiles

The Barndorff-Nielsen approximation only takes single parameters of interest into account, with adjustment for the nuisance parameters. If the combination of several components of a vector of parameters are simultaneously of interest, an approximation is needed, considering a multidimensional parameter space. The same technique as for the single parameters can be used, constructing a multi-parameter profile by constrained optimization of the deviance and then finding adequate tail area approximations, using the observed information functions of these constraint models. Skovgaard (2001) gives an analogous method to construct these profiles, approximating the conditional distribution of the vector of MLEs. Calculations are difficult, as they are computationally very demanding; with increasing number of parameters of interest, even more constraint optimizations around the MLE have to be performed to construct a profile for parameter combinations, giving an adequate resolution. Furthermore, estimates for covariances of likelihood differences and derivatives are needed (Skovgaard, 1996).

These multi-parameter profiles may be of interest to construct confidence regions based on the modified deviance statistic, when the focus lies on the dependence of two or more parameters simultaneously. But presenting a complete confidence region to make inference about model parameters might be a difficult task, both for calculation and interpretation.

3.7 One-sided confidence intervals

Instead of calculating a two-sided quantile to equally segment the given error probability to all tails of the multivariate distribution, it can also be spent exclusively to the upper or the lower tails. Hence, with these quantiles one-sided upper or lower simultaneous confidence intervals can be calculated, setting the opposite limit to $-\infty$ or ∞, respectively.

The use of these one-sided confidence intervals might be arguable, if a further assumption is imposed on a confidence region, that it must comprise a set of parameter values *most strongly supported by the data* (Boyles, 2008). For constructing the one-sided confidence region, an area, which is located in an uninformative region of the parameter space to the experimental question but corresponding to a high likelihood, is 'sacrificed' to the benefit of an area, supporting the one-sided formulated experimental question but comprising a lower likelihood. Therefore, a criticism could be the connection of a predefined hypothesis with the confidence interval, which might be interpreted hypothesis-free; hence, different $(1-\alpha)$ confidence regions might be available for the same dataset, leading to different conclusions according to the imposed requirements on the particular confidence region. Nevertheless, in terms of coverage probability also the one-sided intervals should lead to adequate decisions, and the corresponding test decision has an increased power.

Using a first order approximation to the likelihood by computing confidence intervals based on the profile deviance may have additional benefits in the one-sided case compared to the quadratic approximation. By approximating the likelihood with a quadratic function, the complete likelihood has to be considered. At non equitailed distributions, both tails of the likelihood are adjusting the slope of this quadratic function; that is, the opposite tail might additionally influence the statistic and thereby the confidence limit of interest.

3.8 Accounting for extra dispersion

In most real data settings, e.g. count data will not follow exactly a Poisson process with the assumption of the variance being equal to the mean, or categorical data will not reveal a variance term, arising from the assumption of a binomial distribution. This extra variability in the data may for example be caused by any environmental condition, which has an effect on the observed data, but is not represented by a covariate in the model.

To account for this overdispersion, quasi-likelihood methods can be used (Wedderburn, 1974). Instead of assuming a fixed variance function to construct a likelihood, a loose variance mean dependency can be assumed. The amount of extra variability is then estimated from the data by iteratively assigning weights to each observation, shifting the variability, which cannot be explained by the resulting weighted predictions, into a single, additional dispersion parameter (McCullagh and Nelder, 1989).

An alternative approach is a direct incorporation of an additional dispersion parameter into the variance function of the assumed distribution. This leads to the assumption for example of a negative binomial distribution for count data (Lawless, 1987), and a beta-binomial distribution, or Polya distribution in the multivariate case for categorical data (Agresti, 2007). In opposite to the quasi-likelihood methods, a fixed variance-mean relationship is assumed.

Consequently, the dispersion parameter has to be introduced into the deviance statistic to capture the complete uncertainty about the parameter of interest. A straightforward way is to scale the statistic, exemplary the signed root deviance profile, by

$$\tilde{r}(\beta) = \frac{r(\beta)}{\sqrt{\phi}} \tag{3.30}$$

with ϕ being the dispersion parameter (Venables and Ripley, 2003). At $\phi = 1$ no overdispersion is assumed with increasing ϕ at increasing extra variability.

As a correlation between the dispersion and several other parameters can be expected in models with assumption of distributions featuring variance-mean dependency, simultaneous estimation may get complicated. These difficulties even increase, estimating a profile by constrained optimization.

One approach, which highly simplifies the calculations, is fixing the dispersion parameter

to its MLE over the range of the whole profile. The corresponding confidence region does not take the uncertainty about the dispersion parameter into account; as a consequence, the intervals produced at correlation of the dispersion parameter with the parameter of interest are too short.

As a second approach, the dispersion parameter can be treated as additional nuisance parameter. Estimating the dispersion at each point in the profile is time consuming and a convergence of each optimization is not guaranteed. A high variability of the estimated statistics causes additionally problems for the spline interpolation. Therefore an automation of the confidence interval calculation is complicated, but a controlled computation for specific problems might result in more accurate intervals.

Chapter 4

Simulation study

In most cases it is not possible to cover every relevant situation by a simulation study to characterize or compare several methods. Therefore, only some interesting small sample situations are selected, which might be representative for many practical applications. Furthermore one has to bear in mind, that the representation of the simulation results can also influence the evaluation of the methods, e.g. a compact display by a contour plot might hide detailed information, interpolation or choice of a scale might be misleading; otherwise large tables are rather impossible to survey.

To investigate the characteristics of the profile deviance intervals, simulations for count and categorical data are performed. For categorical data only the simple case of two categories in $k \times 2$ tables is explored; Multinomial distributed data with more than 2 categories are not considered, as corresponding models exhibit a more complex parameterization and more complex optimization algorithms are needed. Starting with the simple case of two sample comparisons should give an idea about the performance of the methods, with no need of considering any adjustments of multiplicity of inferences.

A well established measurement of characterizing the performance of a test is by its size and power. In the multiple testing setting several types of power are available as the any-pair power, average power, etc.. Here, only average power is considered, observing the probability of rejecting at least one hypothesis. A common method of evaluating confidence intervals is by its coverage probability, that is, the probability that the true parameter is included in the interval. The corresponding simultaneous coverage proba-

bility is then defined as the probability of the true parameter vector being located inside the confidence set.

Multiple tests and simultaneous confidence intervals based on the signed root deviance profile (SRDP), based on the quadratic approximation to the likelihood (QA), and based on Barndorff-Nielsen's higher order approximations (HOA) are compared. As the confidence intervals for higher order approximated profiles are only computationally available if at least one event occurs in each group for count and categorical data, in all other cases uninformative intervals with limits $[-\infty; +\infty]$ are passed to the coverage probability calculations. To maintain comparability between these approaches, additionally a modification of the SRDP intervals are calculated, substituting the confidence limits by uninformative ones at observing exclusively zeros in a sample (SRDP 0). Treating the occurrence of parameter estimates at the boundary of the parameter space as special case by calculating SRDP and SRDP0, permits an indirect comparison for the confidence interval performance in the settings.

As the confidence limits are found by projecting the intersections of a profile on the space of a parameter, the profile has to cover the range of the interval. For single practical applications, the computation of a profile can be restricted to an adequate range automatically. A bisection algorithm, which does this in a simple way, is described in Section 5.1. But in order to avoid any influence of these algorithms in a simulation study, the SRDP is calculated by spline interpolation for 100 points in an interval within $[-15; 15]$ on a log or logit scale, covering the complete computationally available parameter space. By this method, computational instabilities at occurrence of groups, where overall zero events were counted, are reduced. For the calculation of p-values, the profiling range can be restricted to the space between zero and the parameter estimates; ± 1 to stabilize the spline interpolation at the margins. Only 25 supporting points are supposed to yield reliable results.

All power simulations are performed only at 1,000 runs for each parameter combination, as the computation of the profiles is quite time consuming. At investigating the behavior at small sample sizes, a large variability for the power estimates can be expected. To maintain comparability of the four methods, at each replication all confidence intervals or tests are based on the same generated dataset. For the simulation of simultaneous

coverage probability a higher resolution of 10,000 replications is chosen, at the cost of investigating only a small number of parameter settings.

4.1 Power and size

A statistical test is based on rejecting a null hypothesis with a given error probability, when interest lies on a specific alternative; hence, the simulation study is designed for investigating parameter settings, clearly distinguishing between the null hypothesis and selected alternatives.

4.1.1 Two-sample comparisons

Count data

Two response variables are generated from a Poisson distribution for each combination of 15 Poisson means in a range between 0.1 and 10. At each simulation step, the two means are estimated in a GLM with Poisson family on the log link. Test statistics and corresponding p-values are calculated for the contrast $c_i = \{-1, 1\}$.

Power and size for two-sided (*left*) and one-sided (*right*) comparisons are shown in Figure 4.1. As the two parameters are exchangeable, only the one-sided test with the alternative hypothesis $H_A : \beta_2 - \beta_1 > 0$ is used.

Under H_0 the QA test holds the nominal level strictly for each setting with a tendency of being quite conservative. The SRDP test shows a more liberal behavior, exceeding the nominal level for some settings, but also showing a significantly higher power than the QA test. When not rejecting H_0 for data with zero counts (SRDP 0) the liberal behaviour under H_0 can be avoided. Also the HOA test does not show this liberal behavior as well as being not as conservative as the QA approach. Furthermore, the differences between the QA and profile based tests in handling zero data becomes noticeable, as for the SRDP power increases up to the boundary of the parameter space. For all other methods, completely zero counts yield in a decrease in power. Correspondingly, these results also apply for one-sided testing.

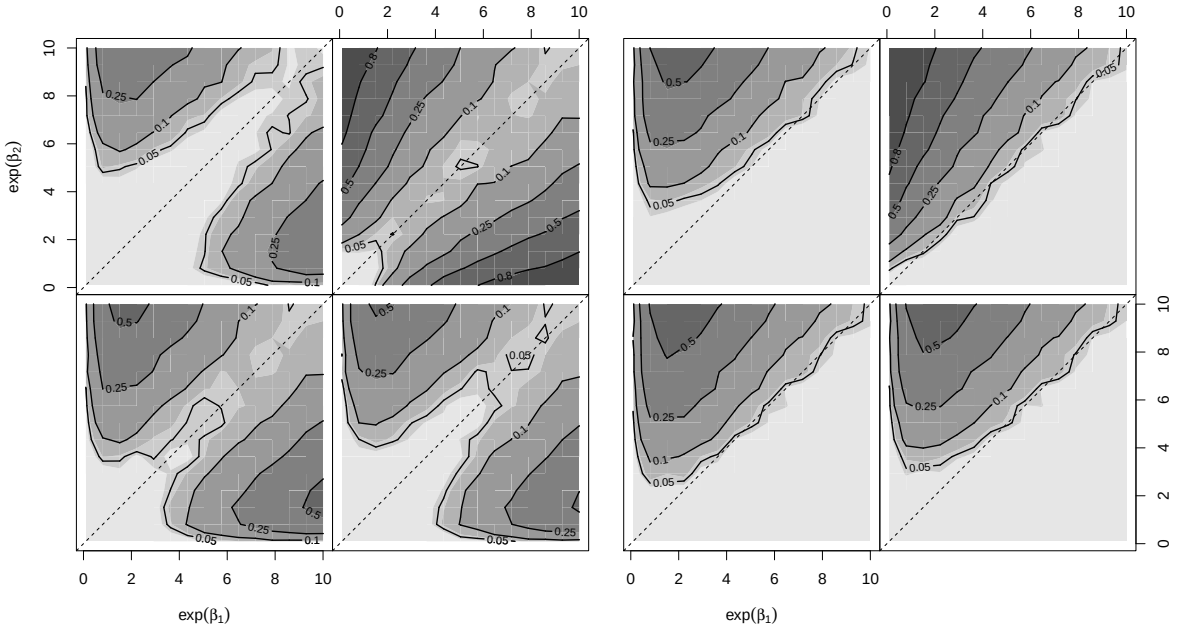

Figure 4.1: Contour plots showing the power and size for profile based tests ($\alpha = 0.05$) for the difference of two Poisson means $\hat{\beta}_1$ and $\hat{\beta}_2$ estimated on the log link. The light gray shaded areas indicate a simulated level below 0.04, and for the next gray level in between $[0.04; 0.06]$; increasing shade shows a higher power. *left*: two-sided comparisons; *right*: one-sided comparisons. Within these panels: *upper left*: QA, *upper right*: SRDP, *lower left*: SRDP0, *lower right*: HOA.

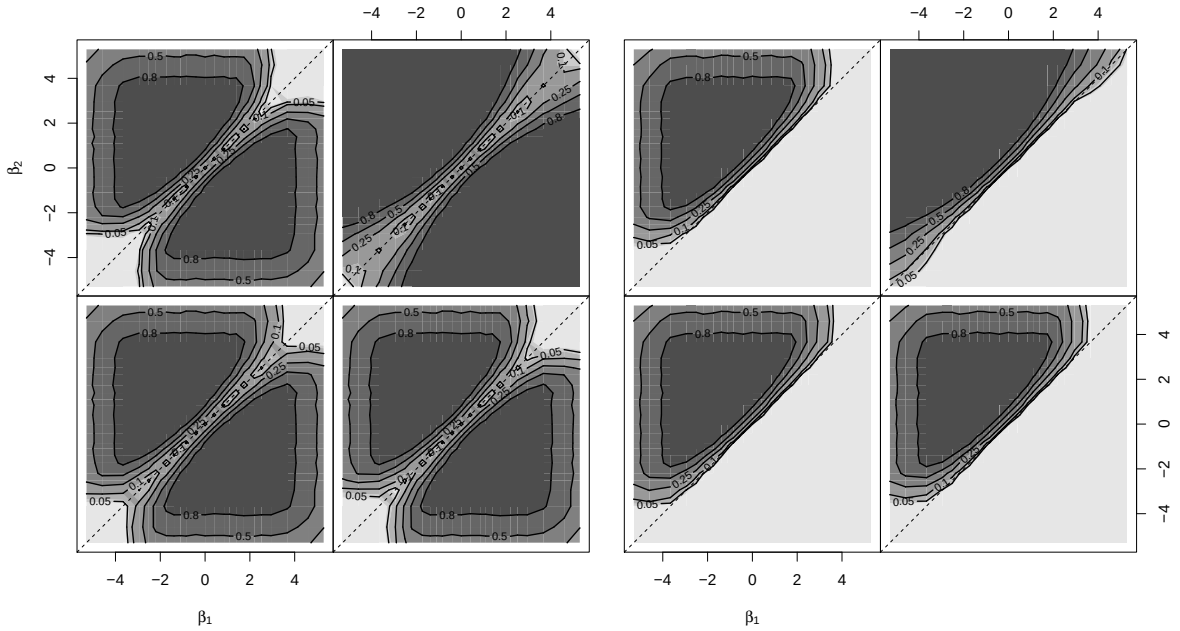

Figure 4.2: Contour plots showing the simulated power and size for profile based tests for the log odds ratio based on the difference of the binomial GLM parameters $\hat{\beta}_1$ and $\hat{\beta}_2$ estimated on the logit link ($\alpha = 0.05$). The light gray shaded areas indicate a simulated level below 0.04, and for the next gray level in between $[0.04; 0.06]$; increasing shade shows a higher power. *left*: two-sided comparisons; *right*: one-sided comparisons. Within these panels: *upper left*: QA, *upper right*: SRDP, *lower left*: SRDP0, *lower right*: HOA.

Categorical data

Data for a 2×2 table is generated from a Binomial distribution for all parameter combinations of 23 values between 0.005 and 0.995, and a population size of 100; The two parameters are exchangeable for two-sided comparisons (Agresti, 2007), resulting in the same performances mirrored at the bisecting diagonal of the parameter space. For each generated table, the probability of successes is estimated on the logit link. Applying the contrast $c_i = \{-1, 1\}$ results in tests for the odds-ratio.

The simulation results for two- and one-sided tests are shown in Figure 4.2. The main focus is laid on small proportions, hence the contour plot axes are shown on the logit scale.

The resolution of observing all combinations of 23 proportions, spread over the whole parameter space, makes it hard to detect any violation of type-I-error rates in the figures. Looking at the raw simulation results, the same outcome as for the count data can be

observed, with the QA approach being conservative, SRDP being liberal, and the HOA method representing a compromise between those two approaches. This becomes also obvious when examining the power and size at very small proportions. Noticeable is also the treatment of counting no event in one of the categories, yielding in a decreasing power at the boundary of the parameter space, except for the SRDP test.

4.1.2 Multiple tests in a one-way layout

To investigate the power and size of multiple tests based on a deviance profile, the simulation setting is extended to the comparison of five samples, as a trade-off between obtaining information about an adequate multiplicity control and computation time. The sample size is set to $n = 1$ per sample.

Representative for some useful multiple comparison procedures, described in Section 3.2.2 and with examples in the Appendix A.2, simulations are performed for many-to-one, Williams-type, and changepoint comparisons. These contrasts show an increasing amount of correlation between the test statistics. At each simulation step, multiple tests are performed, testing the two-sided hypothesis $H_0 : \bigcap_{m=1}^{M} \psi_m = 0$ with $\psi_m = \sum_{j=1}^{k} c_{mj}\beta_j$ for $m = 1, \ldots, M$ local hypotheses. Only a small extract of all combinations of the five sample means β_j are observed. The parameters are partitioned into two groups, each with the same mean, restricting the alternatives to a grid of combinations of only two parameters.

As a direct comparison, Bonferroni adjusted tests are calculated (marked in all figures as lightgray, dotted contour lines).

Count data

As a basis for all count data simulations the five sample means are estimated in a Poisson GLM on the log link. Parameters are chosen from the set of Poisson means $\{0.5, 1, 1.5, 2, 2.5, 5, 7, 10, 15\}$.

For the Dunnett-type comparisons all combinations of the two parameter groups β_1 and $\beta_{2,3,4,5}$ are observed, reproducing a design with one control group and all other treatments in the alternative. Simulated size and power is shown in Figure 4.3 for two-sided and

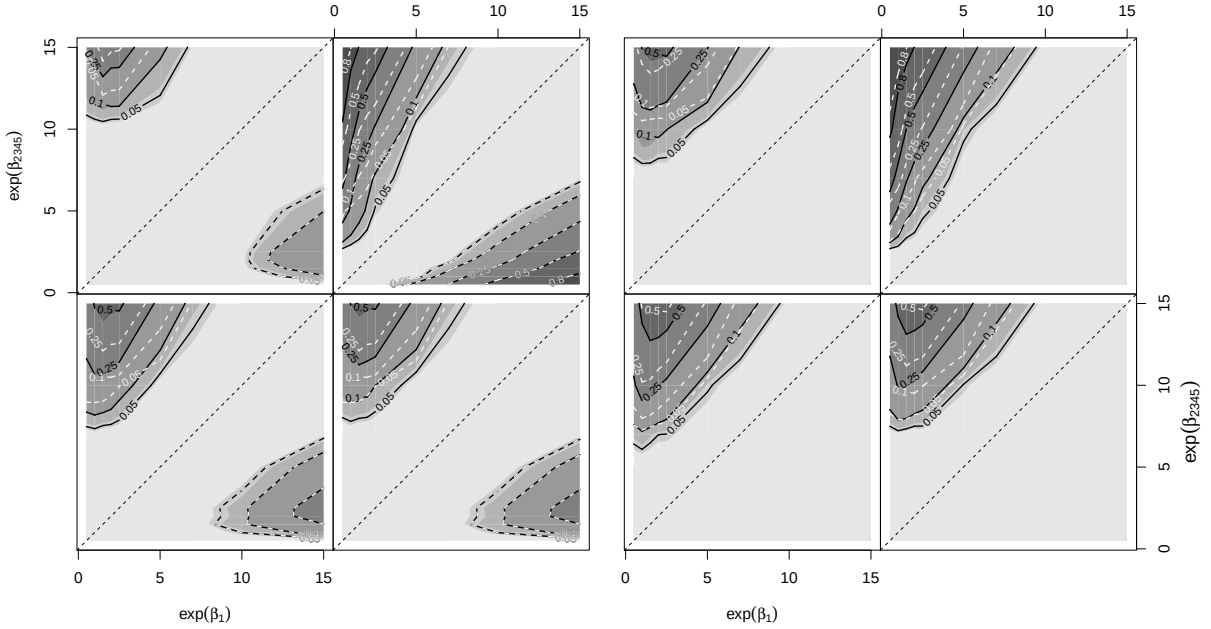

Figure 4.3: Contour plots showing the power and size for profile based tests ($\alpha = 0.05$) for Dunnett-type comparisons of five Poisson means $\hat{\beta}_1$ and $\hat{\beta}_{2,3,4,5}$ estimated on the log link. The lightgray shaded areas indicate a simulated level below 0.04, and for the next grey level in between $[0.04; 0.06]$; increasing shade shows a higher power. *left*: two-sided comparisons; *right*: one-sided comparisons. Within these panels: *upper left*: QA, *upper right*: SRDP, *lower left*: SRDP0, *lower right*: HOA.

one-sided comparisons. The size of the test is for every setting far below the nominal level of 0.05; the performance under the alternative is comparable to the two-sample comparisons for all of the four methods. As the comparisons are performed to a common control group the multiple tests are correlated; therefore the Bonferroni correction yields in a loss in power, assuming uncorrelated tests. A special situation arises at large number of counts in the control group and small ones in the other group. The small number of counts yields in large variance estimates, which are used to estimate the correlation structure for the underlying multivariate normal distribution. In the extreme settings under investigation, this estimation problem dominates, hence the correlation due to the contrasts has virtually no impact. According to this, the Bonferroni correction shows the same characteristics as the 'plugin' method for these settings.

The simulations for the Williams-type contrasts are performed by defining the parameters β_1 and β_5, obtaining the remaining ordered parameters by linear interpolation on the log link. One-sided tests are investigated, assuming that interest lies in a specific direction

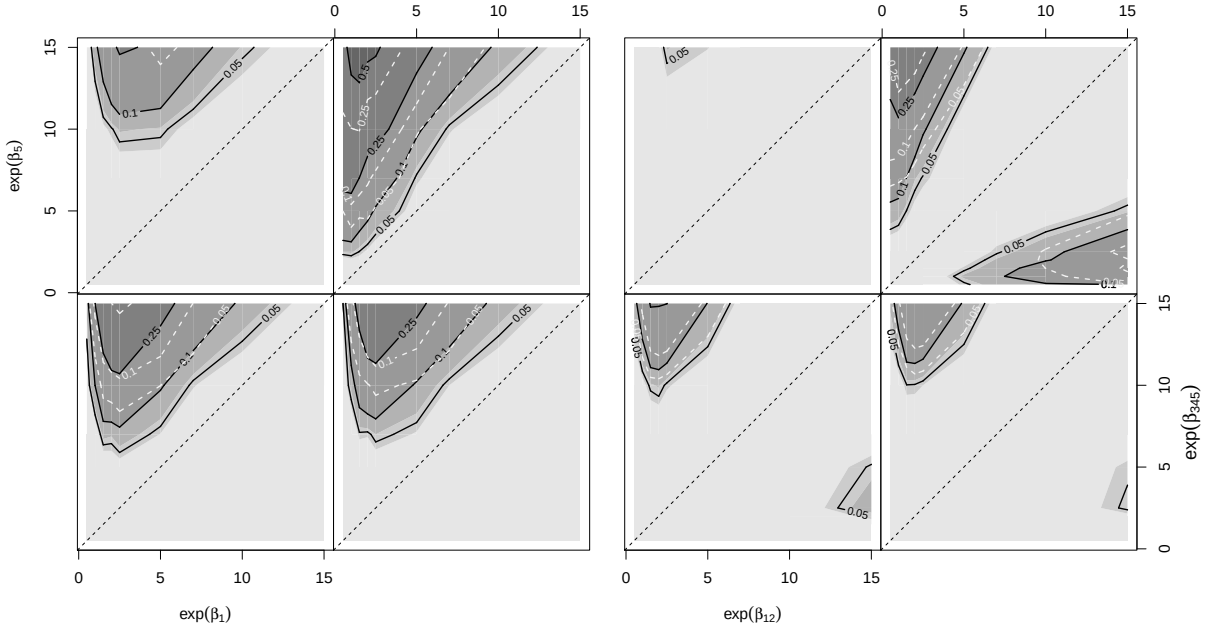

Figure 4.4: Contour plots showing the power and size for profile based tests ($\alpha = 0.05$) for one-sided Williams-type (*left*) and two-sided changepoint (*right*) comparisons of five Poisson means estimated on the log link. The Williams settings are parameterized as β_1 and β_5 assuming a linear trend for the parameters within; for the changepoint contrast, the difference between the parameter groups $\beta_{1,2}$ and $\beta_{3,4,5}$ is observed. The lightgray shaded areas indicate a simulated level below 0.04, and for the next gray level in between $[0.04; 0.06]$; increasing shade shows a higher power. Within the panels: *upper left*: QA, *upper right*: SRDP, *lower left*: SRDP0, *lower right*: HOA.

when testing for a trend. Two-sided tests are used for changepoint detections between the pooled groups $\beta_{1,2}$ and $\beta_{3,4,5}$. The power results, shown in Figure 4.4, are again similar to the ones before. Due to larger correlations between the tests, the difference to the Bonferroni correction increases. The estimation of the correlation in the changepoint setting is not as problematic as in the Dunnett case, as the comparisons are performed between pooled parameter groups and not to a single control group.

Categorical data

Similar to the Poisson settings, simulations are performed for a Binomial GLM on the logit link, applying multiple contrasts on estimates for a $k \times 2$ table. Parameters are chosen from the set of proportions $\{0.005, 0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 0.925, 0.95, 0.975, 0.99, 0.995\}$ with a population size of 100.

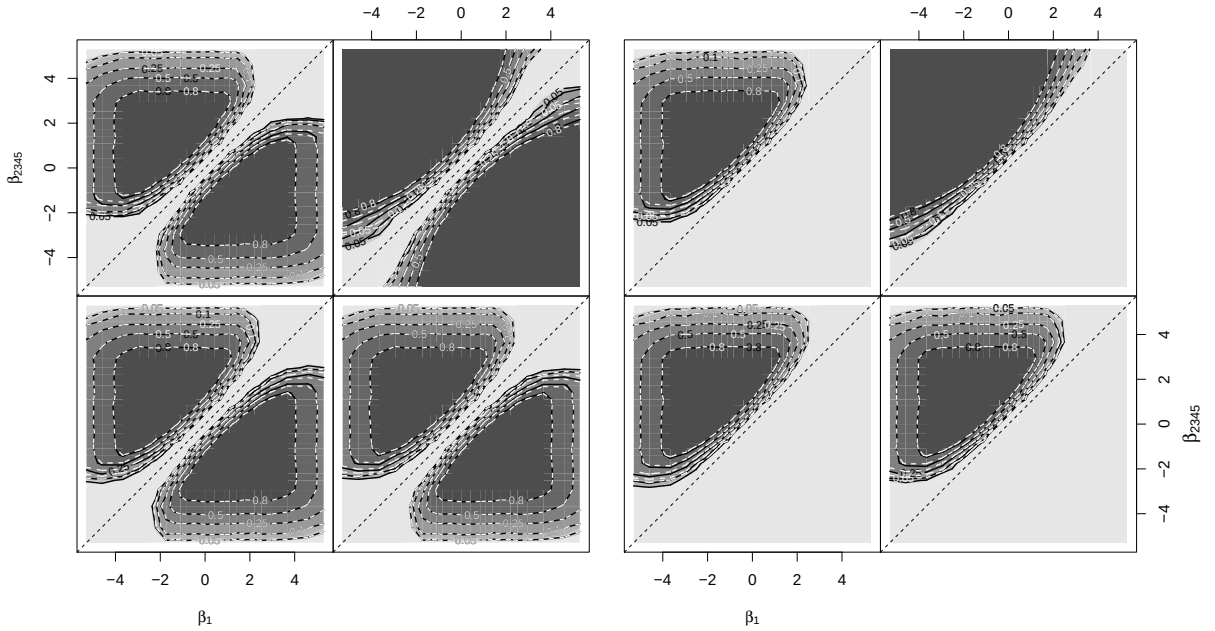

Figure 4.5: Contour plots showing the power and size for profile based tests ($\alpha = 0.05$) for Dunnett-type comparisons of five Binomial proportions $\hat{\beta}_1$ and $\hat{\beta}_{2,3,4,5}$ estimated on the logit link. The lightgray shaded areas indicate a simulated level below 0.04, and for the next gray level in between $[0.04; 0.06]$; increasing shade shows a higher power. *left*: two-sided comparisons; *right*: one-sided comparisons. Within these panels: *upper left*: QA, *upper right*: SRDP, *lower left*: SRDP0, *lower right*: HOA.

For the one- and two-sided Dunnett-type comparisons (Figure 4.5) and for the one-sided Williams-type comparisons (Figure 4.6) the same results apply for the odds ratio comparisons as for the Poisson settings. All four methods hold the nominal level being almost far too conservative and the SRDP tests show an increased power especially for the extreme settings of small proportions. A difference to the Bonferroni adjustment can only be seen at these comparisons of small proportions, where also the correlation estimates have a large impact on a gain in power for the plugin method. It has to be acknowledged, that the presentation of the results for multiple settings over a large range of parameters can not highlight every feature, which might be of interest.

The changepoint comparisons, shown in the right panel of Figure 4.6, exhibit a special situation, where also the SRDP test shows a loss in power at the borders of the parameter space. As for the changepoint contrast mainly pooled parameter combinations are tested, the impact of a single category without any event counted is relatively small. But due to the small sample sizes for these settings, a test might not reject the null hypothesis

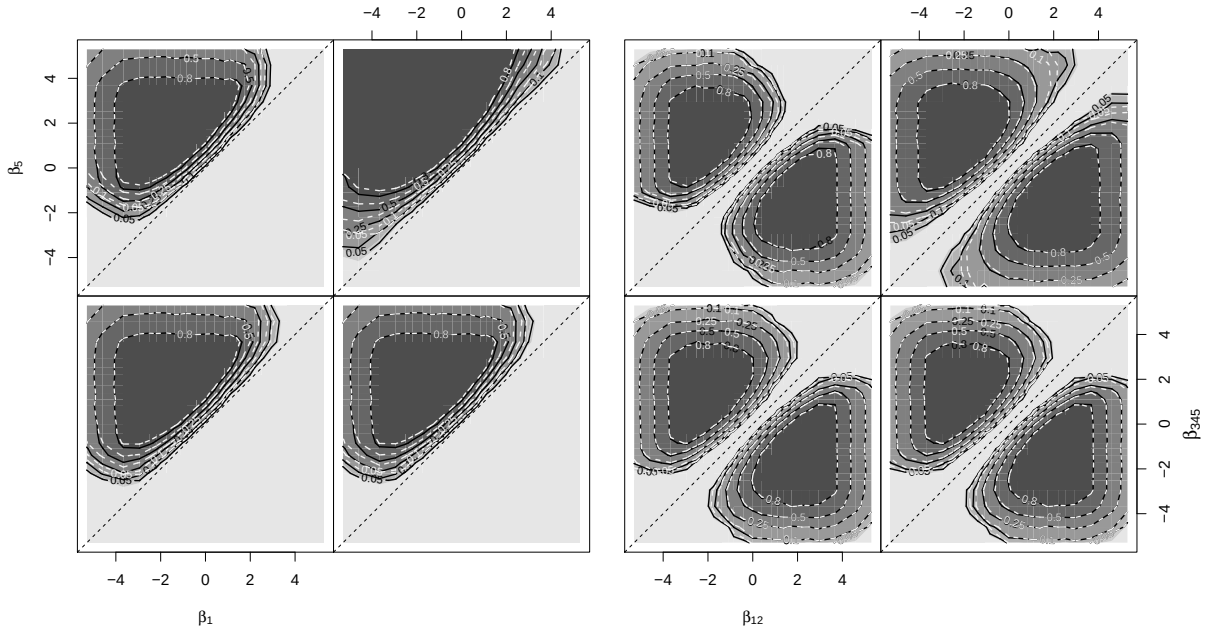

Figure 4.6: Contour plots showing the power and size for profile based tests ($\alpha = 0.05$) for one-sided Williams-type (*left*) and two-sided changepoint (*right*) comparisons of five Binomial proportions estimated on the logit link. The Williams settings are parameterized as β_1 and β_5 assuming a linear trend for the parameters within; for the changepoint contrast, the difference between the parameter groups $\beta_{1,2}$ and $\beta_{3,4,5}$ is observed. The lightgray shaded areas indicate a simulated level below 0.04, and for the next gray level in between $[0.04; 0.06]$; increasing shade shows a higher power. Within the panels: *upper left*: QA, *upper right*: SRDP, *lower left*: SRDP0, *lower right*: HOA.

and leads therefore to conservative results. Furthermore, the calculation of the profiles for changepoint contrasts might be instable as the conditional optimization involves the estimation of all of the five parameters at each step.

4.2 Coverage probability

Confidence intervals are not based on a specific null hypothesis, but should characterize the uncertainty about a parameter estimate with a given error probability. To provide a general view about the application of the presented confidence intervals for a broad range of realistic models, the simulation settings are summarized by parameters of Gamma or Beta distributions, when more than two samples are involved. By varying a location parameter the confidence interval performance at decreasing sample sizes is described. Assuming a fixed dispersion parameter enables the consideration of several realistic pa-

rameter configurations. This allows for a basic representation of coverage probabilities, closer to the practical application than observing only a small set of fixed parameter combinations.

4.2.1 Two-sample comparisons

Count data

Two response variables are generated from a Poisson distribution for each combination of 10 Poisson means in a range between 0.1 and 10. At each parameter combination, the two means are estimated in a GLM with Poisson family on the log link. Then, confidence intervals are calculated for the contrast $c_i = \{-1, 1\}$. The estimated coverage probabilities are presented in Figure 4.7.

The QA intervals are showing an overall conservative behavior, but getting liberal at small mean values in one of the groups. This liberality, in spite of observing a large number of zero event groups, may be caused by adjusting the quadratic profile on the lower tail of the likelihood. Hence, the upper confidence limit is underestimated at largely skewed distributions. Nominal coverage is only reached at larger mean values, which are barely covered by the chosen simulation settings. The SRDP intervals reach the nominal level faster, at smaller mean parameters than the QA method. At the borders of the parameter space the intervals tend to be conservative, but near the border a liberal area can be seen. In this region, the probability of drawing completely zero count data for one of the two samples is high; all information about the variability is then obtained from a single group, yielding in an underestimation of the variance of the difference of the two means. Setting the corresponding interval uninformative for these settings (SRDP0) eliminates this liberality, replacing it with conservative intervals. The HOA method shows a little higher coverage probability than the SRDP0 intervals as a trade-off between the QA and SRDP approach. The one-sided intervals reflect the same results as for the two-sided intervals. Noticeable is the liberal behavior of the HOA approach at one border of the parameter space, resembling the characteristic of the QA intervals for these settings.

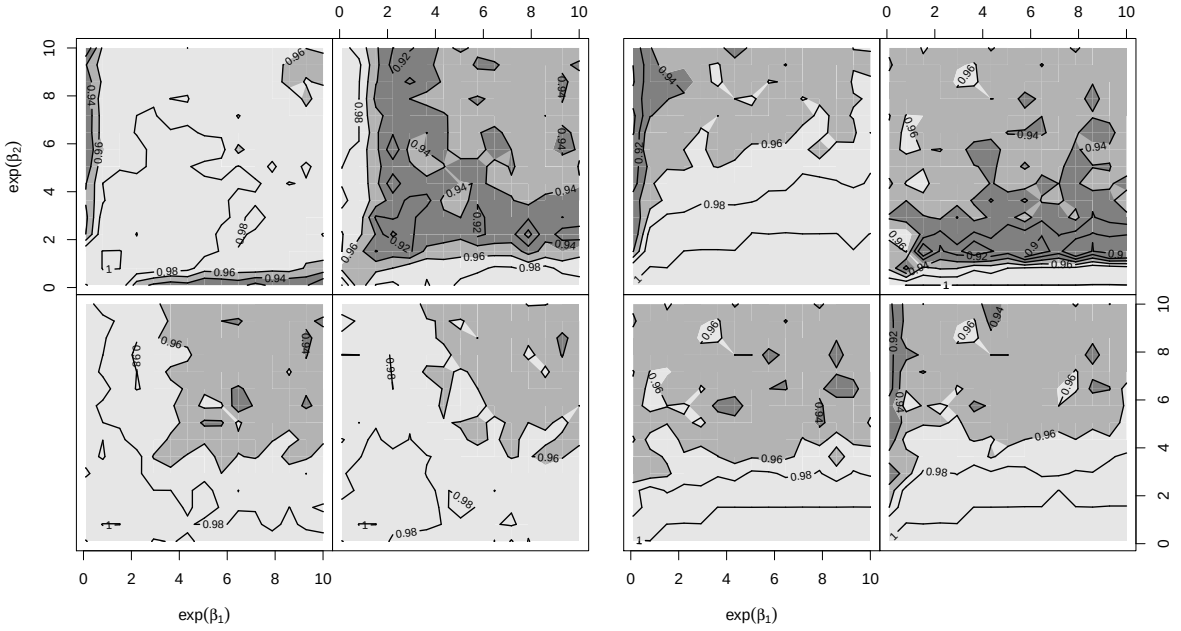

Figure 4.7: Contour plots showing the simulated coverage probabilities for the difference of two Poisson means $\hat{\beta}_1$ and $\hat{\beta}_2$ estimated on the log link ($(1-\alpha) = 0.95$). The light gray shaded areas indicate simulated coverage > 0.96; the dark gray shaded areas < 0.94; with medium gray intensity for $[0.94; 0.96]$; *left*: two-sided comparisons; *right*: one-sided comparisons. Within these panels: *upper left*: QA, *upper right*: SRDP, *lower left*: SRDP0, *lower right*: HOA.

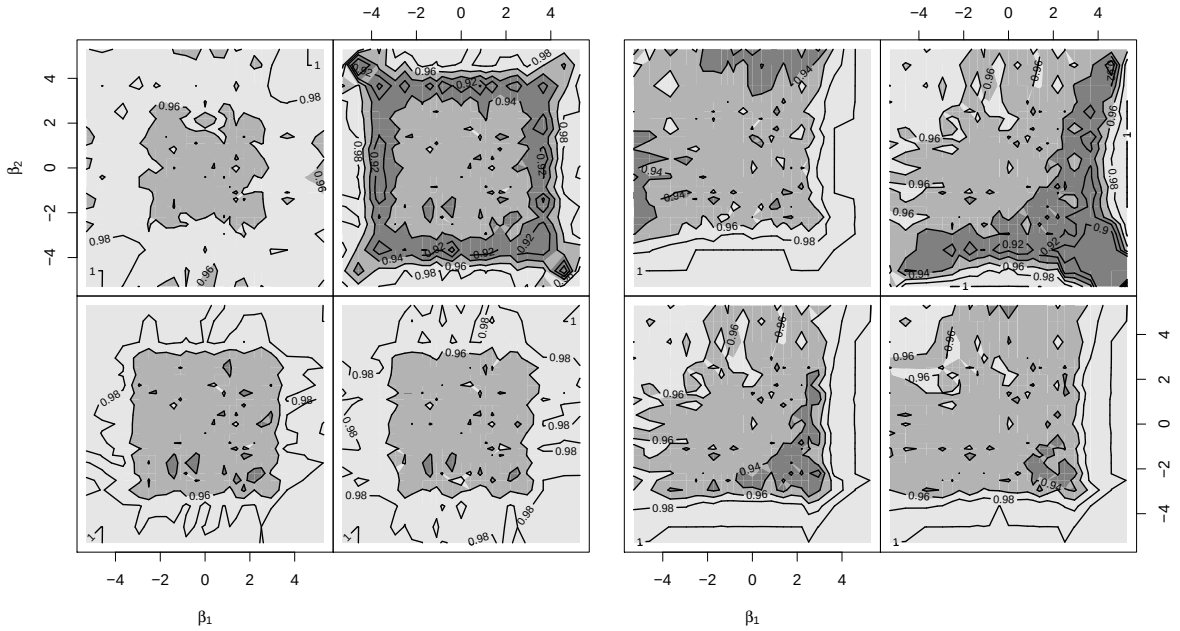

Figure 4.8: Contour plots showing the simulated coverage probabilities for the log odds ratio based on the difference of the binomial GLM parameters $\hat{\beta}_1$ and $\hat{\beta}_2$ estimated on the logit link ($(1-\alpha) = 0.95$). The light gray shaded areas indicate simulated coverage > 0.96; the dark gray shaded areas < 0.94; with medium gray intensity for $[0.94; 0.96]$; *left*: two-sided comparisons; *right*: one-sided comparisons. Within these panels: *upper left*: QA, *upper right*: SRDP, *lower left*: SRDP0, *lower right*: HOA.

Categorical data

Data for a 2×2 table is generated from a Binomial distribution for all parameter combinations of 23 values between 0.005 and 0.995, and a population size of 100. For each generated table, the probability of successes is estimated on the logit link. Applying the contrast $c_i = \{-1, 1\}$ results in confidence intervals for the log odds ratio.

The estimated coverage probabilities are presented in Figure 4.8. The axes are shown on the logit scale to emphasize small proportions.

At the borders of the parameter space all intervals show a conservative behavior, only reaching the nominal level for larger proportions. The area with coverage probability around 0.95 is slightly larger for the SRDP, SRDP0, and HOA intervals in comparison to the QA method. Likewise to the Poisson case, the SRDP intervals are liberal in a region near the boundary due to counting no success or failure in one of the cells of the 2×2 table. This behavior turns conservative for the SRDP0 intervals.

The one-sided QA intervals show a slightly liberal behavior at large odds ratios, when computing lower limits. In comparison to the two-sided intervals, the liberal region of the SRDP method is even increasing; even the SRDP0 intervals are somewhat liberal. Here, the HOA intervals perform best with a large area around the nominal level.

4.2.2 Simultaneous confidence intervals in a one-way layout

To investigate the characteristics of simultaneous confidence intervals based on a deviance profile, the simulation setting is extended to the comparison of five samples, as a trade-off between obtaining information about an adequate multiplicity control and computation time. For Tukey-type comparisons additionally a design with ten groups is used to demonstrate the application to a higher dimensional problem. Replicated data is generated with a balanced sample size of $n_j = 1, 10$ per sample.

Representative for some useful multiple comparison procedures by multiple contrasts, simulations are performed, likewise to the power simulations, for Dunnett-type, Tukey-type, Williams-type, and changepoint comparisons (Appendix A.2). At each simulation step, simultaneous confidence intervals are calculated for the M transformed parameters $\hat{\psi}_m = \sum_{j=1}^{k} c_{mj}\hat{\beta}_j$, with $\hat{\beta}_j$ being the 5(10) sample mean estimates.

As a direct comparison, Bonferroni adjusted simultaneous confidence intervals are calculated.

Count data

As a basis for all count data simulations the five sample means are estimated in a Poisson GLM on the log link.

First, two-sided many-to-one comparisons are performed, using a Dunnett-type contrast matrix. The responses are generated from a Poisson distribution with parameters being drawn from a Gamma distribution with density

$$f(\beta) = \frac{\beta^{a-1} e^{-\frac{\beta}{s}}}{s^a \Gamma(a)} \tag{4.1}$$

with $a > 0$ being a shape and $s > 0$ a scale parameter. Mean and variance are $E(\beta) = a$ and $Var(\beta) = as^2$ (Johnson et al., 1994). Several parameter settings are investigated,

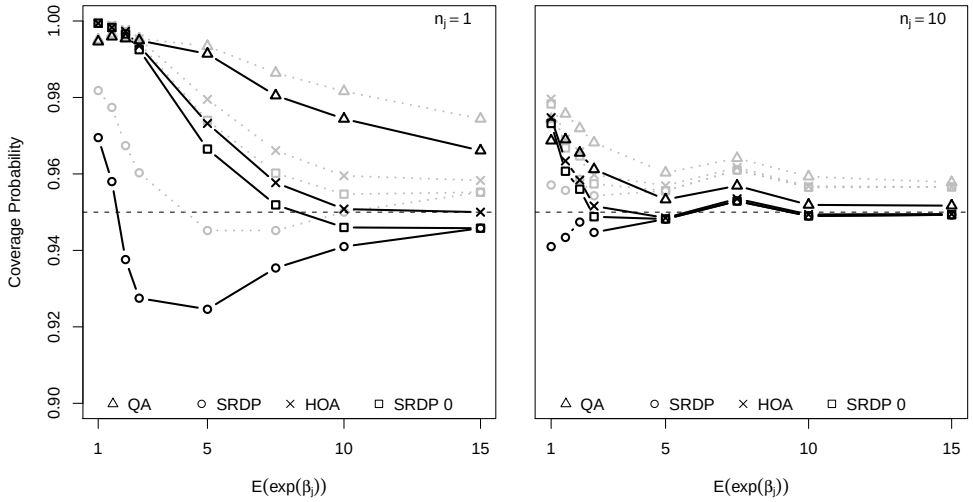

Figure 4.9: Dunnett-type comparisons for 5 means estimated in a Poisson GLM. Simulation settings are varied by the mean of a Gamma distribution $E(\beta_j)$ on the logarithmic scale and the number of replicates per group n_j. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *gray lines* represent multiplicity adjustment by Bonferroni.

fixing the scale parameter at $s = 1$ and varying the mean of the Gamma distribution for 8 locations between 1 and 15.

The results for two-sided Dunnett comparisons are shown in Figure 4.9. All methods are conservative at a small number of counts, but will converge to the nominal level with increasing sample sizes. The QA method shows the most conservative behavior, even showing some conservativeness at larger sample sizes. The SRDP method first yields some liberal decisions before approaching the nominal level with increasing sample size. This effect is caused by groups with zero counts and will disappear when omitting these settings by the SRDP 0 method. Overall the first order approximations (SRDP, SRDP0) are a bit liberal. The higher order approximation (HOA) shows the best results, converging fast and directly to the nominal level.

In many applications, comparisons to a control are needed with interest in only a one-sided alternative. Therefore, the simulation is repeated for one-sided confidence intervals, as discussed in Section 3.7. As a directed alternative is of interest, the parameter settings are restricted to this one-sided problem; the sample means, generated by a Gamma distribution are ordered, with the control group having the smallest value. Hence, the results of the one-sided versus the two-sided simulations are not directly comparable, generating

smaller counts for the one-sided control. Especially for Dunnett-type contrasts, this has an enormous impact, as the control group is used in all of the local comparisons. The results are presented in Figure A.1 in the Appendix. The one-sided intervals show quite the same characteristics as the two-sided ones; as expected the QA, SRDP0, and HOA intervals behave more conservative, and the SRDP is more liberal at small mean values due to the ordered means settings.

If the alternative should be restricted even further, a trend may be detected for the ordered samples by a Williams contrast. Only in rare situations two-sided alternatives are of interest for this contrast, thus the simulation is conducted for one-sided confidence intervals detecting an increasing trend. Although the Williams-type contrast allows for multiple convex response shapes, only a linear trend is considered for the simulations. This is realized by generating the minimum and maximum sample means by a Gamma distribution and finding the means for the intermediate groups by linear interpolation. The simulated coverage probabilities are visualized in Figure 4.10. No much difference can be detected compared to the Dunnett-type simulations. Even for the more complicated correlation structure the nominal level is reached at 'high' sample sizes; the deviance profile methods, especially the HOA approach, shows a faster convergence to this level with increasing sample sizes.

To investigate the interval characteristics at a large number of comparisons and a larger number of parameters, a simulation for Tukey-type comparisons between 5 and 10 group means are conducted (Figure 4.11). In both cases the quadratic approximation is far to conservative at these small samples. The profile methods are converging faster to the nominal level at increasing sample sizes. With an increased number of parameters to compare, all methods become a bit more conservative. This conservativeness can be expected, as the increased number of comparisons $((k(k-1))/2)$ will not be obtained for free.

At last, coverage probabilities of two-sided confidence intervals for changepoint contrasts are investigated. The parameters of interest are highly correlated, as each is a representation of a slightly different linear combination of all original model parameters. Hence, this contrast represents also a challenge for the conditional optimization algorithm. The sample means are generated from a Gamma distribution without any further restriction

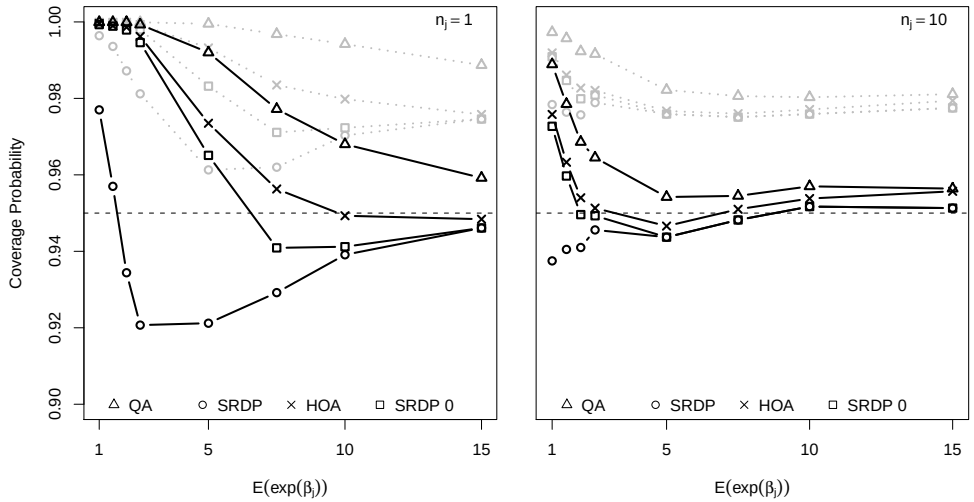

Figure 4.10: One-sided Williams-type comparisons for 5 ordered means estimated in a Poisson GLM assuming a linear trend. Simulation settings are varied by the mean of a Gamma distribution $E(\beta_j)$ on the logarithmic scale and the number of replicates per group n_j. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *gray lines* represent multiplicity adjustment by Bonferroni.

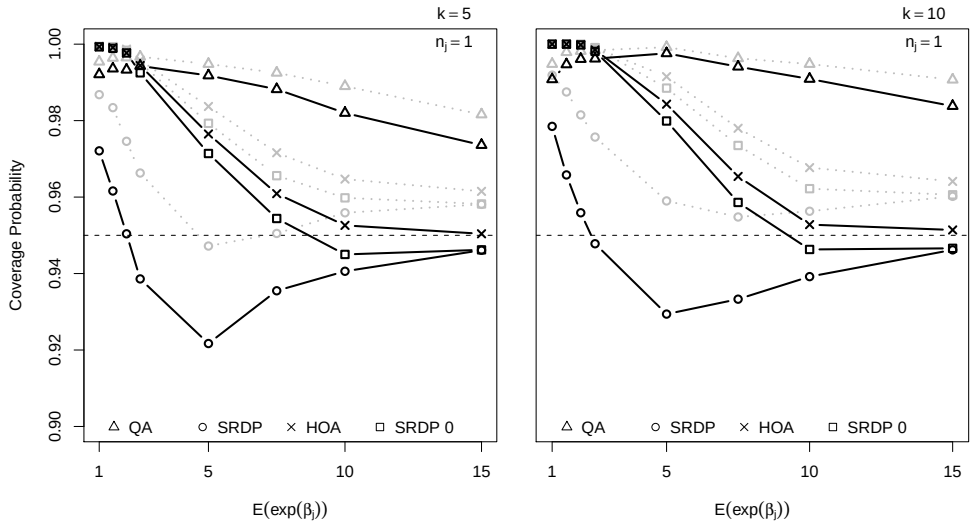

Figure 4.11: Tukey-type comparisons for 5 (*left*) and 10 means (*right*) estimated in a Poisson GLM. Simulation settings are varied by the mean of a Gamma distribution $E(\beta_j)$ on the logarithmic scale and the number of replicates per group n_j. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *gray lines* represent multiplicity adjustment by Bonferroni.

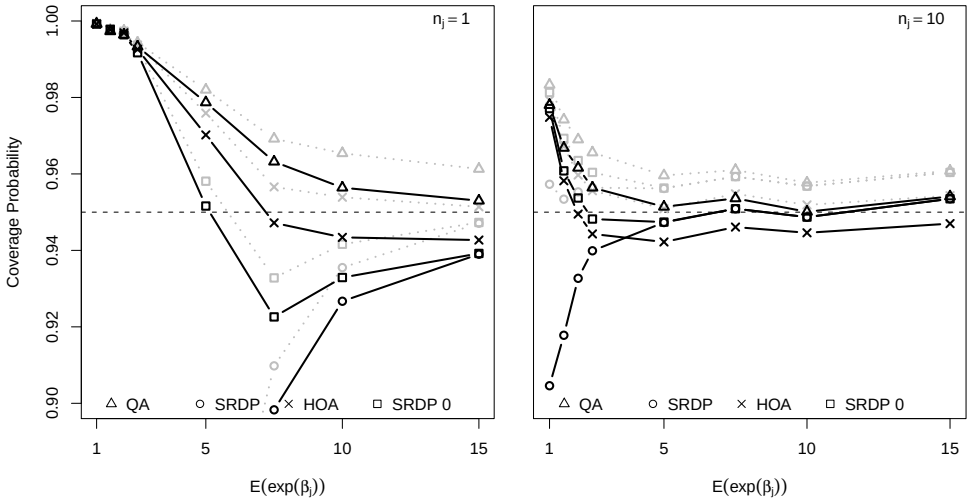

Figure 4.12: Changepoint comparisons for 5 means estimated in a Poisson GLM. Simulation settings are varied by the mean of a Gamma distribution $E(\beta_j)$ on the logarithmic scale. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *gray lines* represent multiplicity adjustment by Bonferroni.

similar to the two-sided Dunnett simulation. The results are found in Figure 4.12.

Due to the large amount of zero counts at small sample means, the SRDP is liberal; a change to larger intervals at small sample sizes can not be found for the chosen simulation settings. Numerical problems may occur for the conditional parameter optimization and spline interpolations; the number of supporting points to construct a profile has to be increased to obtain accurate simulation results. Some numerical instabilities may cause the SRDP 0 and especially the HOA method to underestimate the nominal level.

Categorical data

The observations in a 5×2, (or 10×2) table are generated from a Binomial distribution with sampled population size of 100 for each group. The parameters of the Binomial distribution are obtained from a Beta distribution

$$f(\beta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \beta^{a-1}(1-\beta)^{b-1} \tag{4.2}$$

with $a > 0$ and $b > 0$ being two shape parameters (Johnson et al., 1995). Mean and variance are $E(\beta) = \frac{a}{a+b}$ and $Var(\beta) = \frac{ab}{(a+b)^2(a+b+1)}$. To simplify the parameterization, a dispersion parameter $\phi = \frac{1}{a+b}$ is introduced and fixed at $\phi = 0.01$, varying the expectation

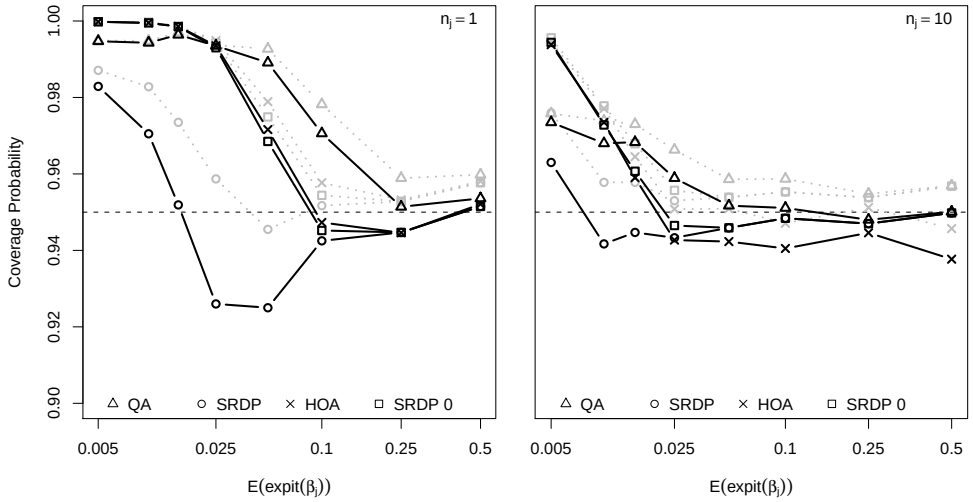

Figure 4.13: Dunnett-type comparisons for 5 parameters estimated in a Binomial GLM. Simulation settings are varied by the mean of a Beta distribution $E(\beta_j)$ on the logit link and the number of replicates per group n_j. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *gray lines* represent multiplicity adjustment by Bonferroni.

over all group means in each simulation setting for 8 values between 0.005 and 0.5. At each simulation step the group means are estimated on the logit link in a Binomial GLM. Simultaneous confidence intervals for the odds-ratio are calculated, using the same types of contrasts as for the count data simulations.

As the simplest setting, two-sided many-to-one comparisons are performed. The simulation results are presented in Figure 4.13.

The same characteristics can be found as for the count data simulations; the QA approach is most of the time conservative, whereas the SRDP, SRDP0, and HOA method are converging faster to the nominal level than the QA method. The HOA approach is liberal at large sample sizes; this might also be caused by numerical instabilities, obtaining less liberal results, when increasing the number of supporting points for the conditional optimization and spline interpolation. This problem is enlarged for contrasts exhibiting a more complex correlation structure.

Just as for the count data, also one-sided Dunnett-type comparisons are performed for the odds-ratio. Again, the result presented in Figure A.2 in the Appendix, are comparable to the two-sided ones. The simulations show clearly a superiority of the profile method

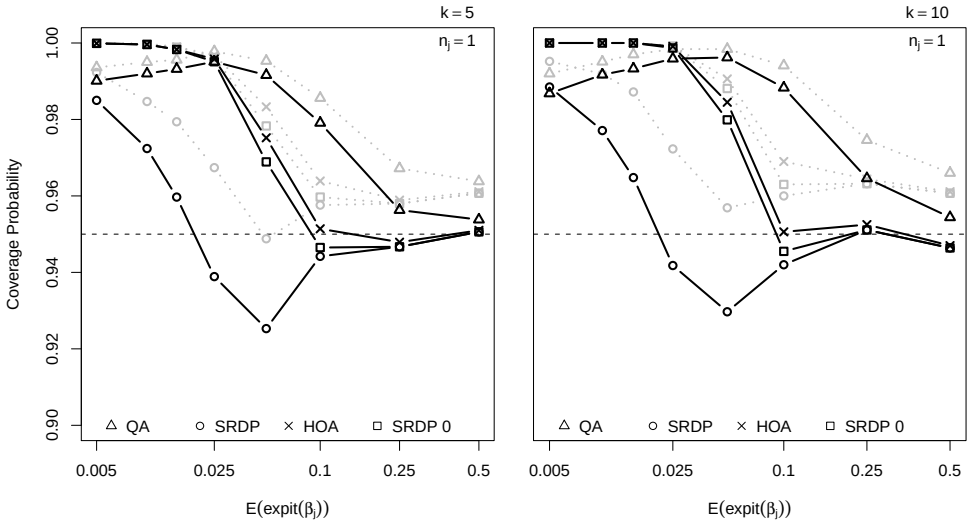

Figure 4.14: Tukey-type comparisons for 5 (*left*) and 10 proportions (*right*) estimated in a Binomial GLM. Simulation settings are varied by the mean of a Gamma distribution $E(\beta_j)$ on the logit scale. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *grey lines* represent multiplicity adjustment by Bonferroni.

compared to the quadratic approximation in terms of coverage probability. Additionally investigating a systematic bias at detailed length, which is a known problem for confidence intervals for the odds ratio (Staicu, 2009), would require far more simulation settings over a wider range of parameter combinations and is therefore omitted here.

Coverage probabilities for one-sided Williams-type contrasts and two-sided changepoint comparisons are shown in Figure A.3 and Figure A.4 in the Appendix. Likewise to the Poisson setting only linear trends are assumed for the Williams-type confidence intervals. The minimum and maximum sample proportions are generated from a Beta distribution, interpolating the intermediate group means by a logistic function. The changepoint settings are chosen similar to the parameter layout for the two-sided many-to-one comparisons. Both results resemble the outcome of the Poisson simulations, showing a superiority of the profile deviance methods compared to the QA approach. Again the HOA method shows computational instabilities at large sample sizes yielding in an increased liberal performance at large sample sizes and expected proportions around 0.5 respectively.

The possibility of performing a larger number of comparisons and comparing a larger number of groups is shown exemplary for Tukey-type comparisons in Figure 4.14. Again the quadratic approximation is highly conservative, whereas the profile methods reach the

nominal level earlier when increasing the probability of observing a success.

4.2.3 Overdispersion

The simulation settings presented in this section are comparable to the ones in Section 4.2.2, with introducing additional variation for each observation. As the number of replications to estimate each parameter is the key factor in estimating any overdispersion parameter, only three different sample mean configurations are used to generate the five group means, but a range of 8 different sample sizes per group n_j from 2 to 20 are included instead. Settings with only one observation per group certainly do not allow estimating any overdispersion. Only two-sided Dunnett- and one-sided Williams-type comparisons are performed.

Count data

To add some amount of extra dispersion to the responses obtained from a Poisson distribution, the expected value for each observation is generated from a Gamma distribution (Eq. 4.1) with shape parameters $a_i = \exp(\sum_{j=1}^{k} x_{ij}\beta_j)$, $i = 1, \ldots, N$, and scale parameter $s = 1$. The resulting Gamma-Poisson distribution should represent some fairly overdispersed count data, which may occur in many real data applications. At each simulation step, parameters are estimated on the log link assuming a quasipoisson model (Section 3.8) conditional on the estimated overdispersion parameter.

The resulting coverage probabilities for Dunnett- and Williams-type comparisons are shown in Figure 4.15 and Figure 4.16.

The results depend on three different factors: the mean of the Gamma-Poisson distribution, the sample size per group, and the amount of overdispersion. The dominating effect here is caused by underestimating the amount of overdispersion at small sample sizes as the profiles are only scaled by the dispersion parameter estimate (Eq. 3.30). This leads to a liberal performance for all of the presented methods. Especially at high mean values and small sample sizes the amount of extra variability is underestimated. At small mean values, the effect of zero counts gain more weight, leading to a conservative behavior of all methods except for the SRDP intervals in the range of evaluated settings. This can

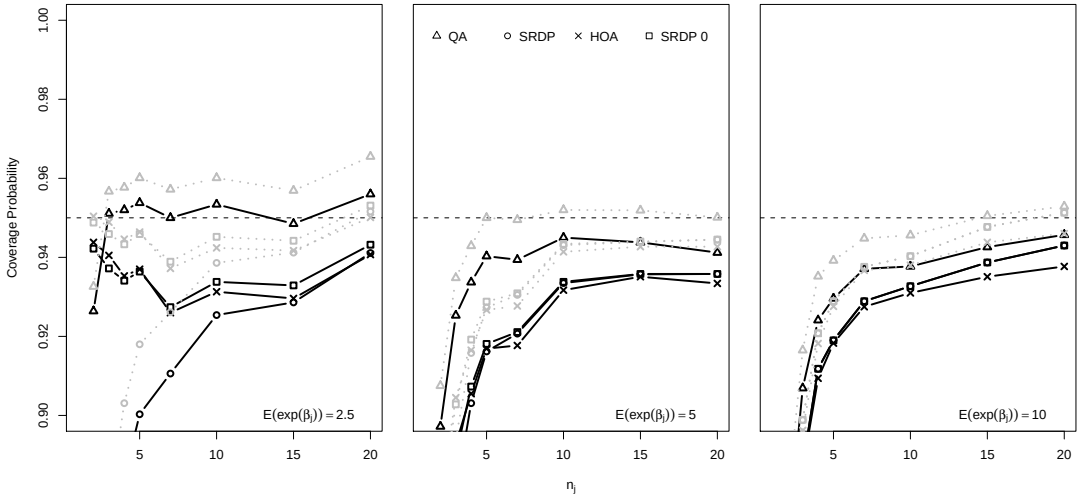

Figure 4.15: Dunnett-type comparisons for 5 means estimated in a quasipoisson GLM. Simulation settings are varied by the mean of a Gamma distribution $E(exp(\beta_j))$ on the logarithmic scale and the number of replicates per group n_j. Fairly overdispersed count data is generated from a Gamma-Poisson distribution at each simulation step. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *gray lines* represent multiplicity adjustment by Bonferroni.

nicely be seen for the Williams-type intervals, where the imposed order restriction on the parameter settings leads to a higher number of zero counts in the 'control' group.

Categorical data

Additional variation is introduced to the categorical data by generating the expectations for each proportion from a Beta distribution (Eq. 4.2) with $\phi = 0.01$ and $p_i = expit(\sum_{j=1}^{k} x_{ij}\beta_j)$, $i = 1, \ldots, N$. Like in the count data setting, counts of successes and failures generated from this mixture of a Beta and Binomial distributions should represent actual fairly overdispersed Binomial data. For each generated dataset confidence intervals based on linear combinations of parameters estimated in a quasibinomial model (Section 3.8) are calculated.

Simulation results for Dunnett- and Williams-type contrasts are shown in Figure 4.17 and Figure 4.18. The same problems as for the Poisson distributed data gets obvious, underestimating the extra variability at small samples, leading to liberal decisions. Additionally, the large number of zeros at small proportions counteracts the liberal behavior of the QA, HOA, and SRDP0 intervals.

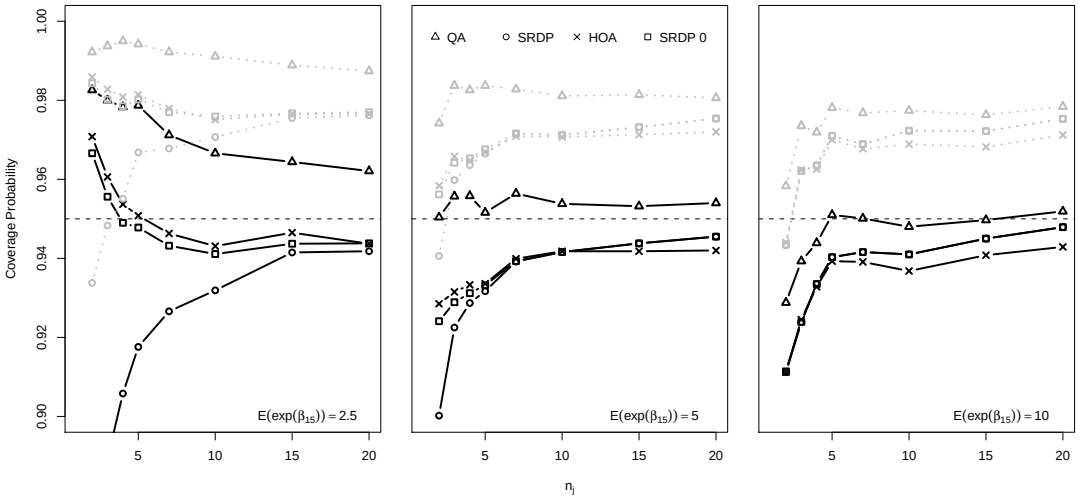

Figure 4.16: One-sided Williams-type comparisons for 5 ordered means estimated in a quasipoisson GLM. Simulation settings are varied by the mean of a Gamma distribution $E(\beta_j)$ on the logarithmic scale and the number of replicates per group n_j. Fairly overdispersed count data is generated from a Gamma-Poisson distribution at each simulation step assuming a linear trend. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *gray lines* represent multiplicity adjustment by Bonferroni.

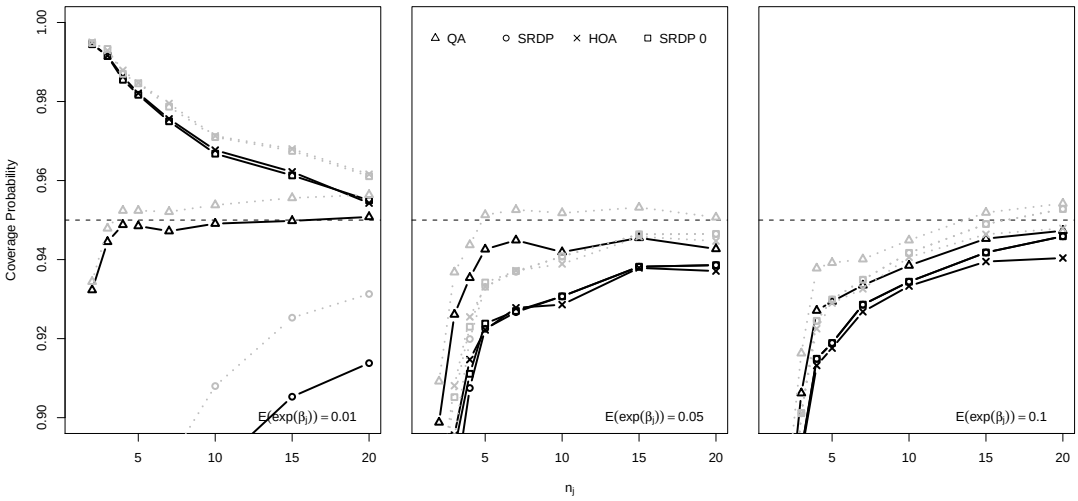

Figure 4.17: Dunnett-type comparisons for 5 parameters estimated in a quasibinomial GLM. Simulation settings are varied by the mean of a Beta distribution $E(\beta_j)$ on the logit link and the number of replicates per group n_j. Overdispersion is introduced by generating data from a Beta-Binomial distribution. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *gray lines* represent multiplicity adjustment by Bonferroni.

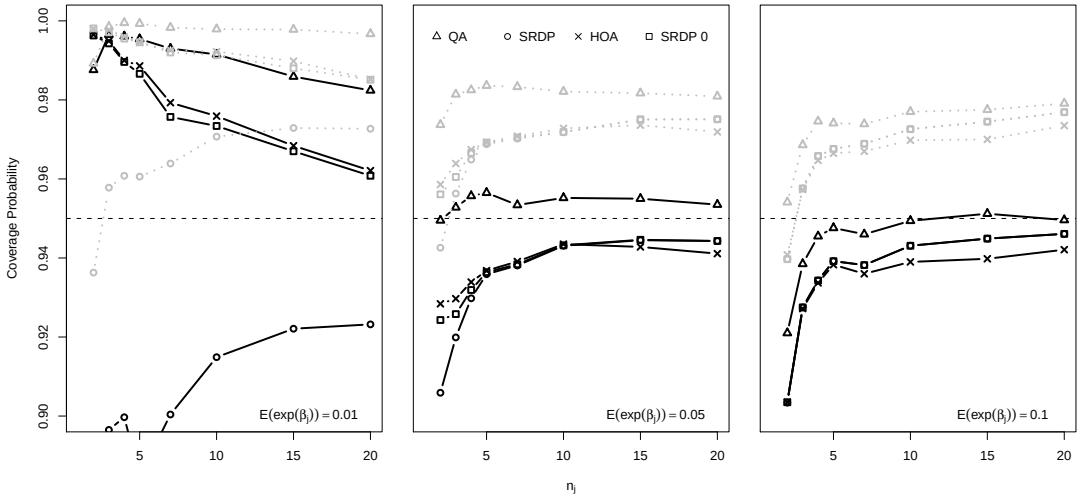

Figure 4.18: One-sided Williams-type comparisons for 5 ordered parameters estimated in a quasibinomial GLM. Simulation settings are varied by the mean of a Beta distribution $E(\beta_j)$ on the logit link and the number of replicates per group n_j assuming a linear trend. Overdispersion is introduced by generating data from a Beta-Binomial distribution. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *gray lines* represent multiplicity adjustment by Bonferroni.

Chapter 5

Example evaluation by implemented software

5.1 Computational Issues and Software Implementation

An R (R Development Core Team, 2009) package `mcprofile` is provided, enabling the computation of signed root deviance profiles for linear transformations of GLM model parameters. The profiles can be adjusted by Barndorff-Nielsen's higher order approximation. Based on the resulting `mcprofile` objects, simultaneous confidence intervals and multiple tests can be calculated; additionally several graphical visualizations of the profiles are provided.

Given an object of class `glm` or `lm`, and a contrast matrix with the same number of columns as the number of parameters in the model, signed root deviance profiles are calculated for a user-defined number of points around the MLE. A matrix with minimum and maximum limits can be provided to define the parameter space for each profile; if limits are omitted, they are determined automatically by a bisection algorithm, searching for projections, where the statistic is approximately at the value of 1.5 times the quantile of a standard Normal distribution at a Bonferroni adjusted α-level. For the conditional optimization the function `optim` is used, applying a quasi-Newton algorithm, minimizing the deviance statistic with an restriction on the model parameters. An extract of the function is shown

in the Appendix (p. IV). The output of the function is an object of class `mcprofile` containing amongst others a list with a sequence of parameter values around the MLE of interest, the corresponding signed root deviance statistic, and the conditional parameter estimates of the original model parameterization. Additionally, functions are provided for interpolation, based on the R function `splinefun`.

The higher order approximations are based on modified R code of Brazzale et al. (2007). After modifying the design matrix of the model, a conditional optimization is performed with the R function `glm.fit` (Venables and Ripley, 2003), using the predictions, obtained by the conditional estimation in the `mcprofile` object, as offset. The iteratively reweighted least square estimation provides variance-covariance estimates for the full model and the conditional fit, which can be used to obtain the observed Fisher information function for both models, after multiplication with an adequate contrast matrix. A simplified extract of the R code is shown in the Appendix (p. V).

Simultaneous confidence intervals and multiple tests are constructed by the functions `confint` and `test`. The associated p-values and quantiles of a multivariate normal distribution are calculated with help of the R package `mvtnorm` (Genz et al., 2009). Extracts of the R code providing statistical inference can be found in the Appendix (p. VI).

An important issue about the software implementation is the computation time. The time consumption will of course increase with the number of conditional estimations representing the profile, and the number of parameters in the model. But also the distributional assumption plays a role; for Gaussian distributed data fast convergence is achieved, whereas for highly skewed distributions each conditional parameter estimation takes more time. At the boundary of the parameter space the computational time is maximal, as per default the profile is calculated over a broad space with large difficulties of reaching convergence at all. At last, the number of non-zero coefficients in the contrast matrix has a high impact on computation time. Differences of single parameters, as for Dunnett-type contrasts, are easily profiled; the parameters of interest can in this case directly be represented by an adequate design matrix, performing the constrained optimization with an iteratively reweighted least square algorithm. Otherwise, for comparing pooled groups of parameters, like for Williams-type and changepoint contrasts, more time is needed due to a more complex parameter optimization. For other user-defined contrast matrices the

gradient function of the deviance has to be found numerically during each optimization step; this increases computation time and may cause a larger variability in the results.

5.2 Evaluation of the examples

As all examples represent dose-response studies in a one-way layout with a negative control, the minimum effective dose is detected by calculating simultaneous confidence intervals for Dunnett-type contrasts. Additionally, confidence intervals based on Williams-type contrasts are constructed, with $C1$ being the comparisons of the last with the first parameter, pooling the neighbor of the last parameter for each further contrast.

Exemplary R code is given to demonstrate the application of the implemented software. (It is assumed, that the datasets are imported into `R` and additional packages have been loaded by `library(mcprofile)` and `library(multcomp)`.)

5.2.1 Angina

For the angina example (Section 2.1) a normal distributed response is assumed. Therefore, the quadratic approximation exactly matches the deviance profiles. $(1 - \alpha) = 0.95$ simultaneous confidence intervals are calculated, assuming a t-distribution with 45 residual degrees of freedom.

```
> str(angina)
'data.frame': 50 obs. of 2 variables:
 $ dose    : Factor w/ 5 levels "control","1",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ response: num  12 19.1 14.2 11.2 16.2 ...
```

First, a linear model is fit to the data, estimating the group means by omitting the default parameterization of a common intercept.

```
> fit <- lm(response ~ dose-1, data=angina)
```

To construct confidence intervals for the parameters of interest, a contrast matrix has to be specified, suitable for the vector of estimated model parameters. This is done by a user-friendly function provided by the package `multcomp`.

```
> K <- contrMat(table(angina$dose), type="Dunnett")
```

The signed root deviance profiles can be calculated by

```
> mcp <- mcprofile(object = fit, CM = K)
```

As default, 100 supporting points are used to estimate the profile over a range, estimated by a bisecting method. Plotting the profiles by

```
> plot(mcp)
```

displays them as straight lines, which can be expected for a Gaussian linear model. Simultaneous lower confidence limits are calculated by

```
> ci <- confint(mcp, alternative="greater")
```

shown in Table 5.1. p-values of a corresponding hypotheses test are computed by

```
> test(mcp, alternative="greater")
```

Williams contrasts are applied in the same way, substituting the contrast matrix simply by

```
> K <- contrMat(table(angina$dose), type="Williams")
```

before calculating profiles, confidence intervals, and tests.

Table 5.1: Lower confidence limits for many-to-one comparisons for the angina data example.

Comparison	Estimate	Lower limits
Dose_1 - Control	2.10	-1.35
Dose_2 - Control	3.40	-0.05
Dose_3 - Control	5.00	1.55
Dose_4 - Control	10.50	7.06

The lower confidence intervals indicate that dose 3 is the lowest dose showing a significant difference to the control. A significant trend is detected by the confidence intervals for Williams-type contrasts given in Table 5.2. The single Williams contrast C 1 is the same as the comparison of Dose_4 - Control in the Dunnett contrast, but yielding different lower confidence limits. This effect is caused by a higher correlation between the Williams comparisons, producing a higher value for the critical value computed from the multivariate normal distribution.

Table 5.2: Lower confidence limits for Williams-type trend detection for the angina data example.

Comparison	Estimate	Lower limits
C 1	10.50	7.44
C 2	7.75	5.09
C 3	6.30	3.80
C 4	5.25	2.82

5.2.2 Micronucleus assay

As the micronucleus example (Section 2.2) consists of replicated count data, a quasipoisson model is assumed, accounting for overdispersion within all replications for each dose group. The estimated dispersion parameter is $\hat{\phi} = 0.94$; hence the variability within a dose groups seems to be a bit smaller than assumed under a Poisson distribution.

The Dunnett-type confidence intervals of Table 5.3 are calculated by

```
> str(Mutagenicity)
'data.frame': 31 obs. of  2 variables:
 $ Treatment: Factor w/ 6 levels "Vehicle","Hydro30",..: 1 1 1 1 1 ...
 $ MN       : int  1 2 2 2 3 3 5 2 4 4 ...
> fit <- glm(MN ~ Treatment-1, data=Mutagenicity,
+                 family=quasipoisson(link="log"))
```

The SRDP intervals are computed as for the angina example in Section 5.2.1; all comparisons are performed to the solvent control, omitting the positive control for simplicity.

```
> K <- contrMat(table(Mutagenicity$Treatment), type="Dunnett")
> K2 <- K[-5,]
> mcp <- mcprofile(fit, K2)
> ci <- confint(mcp, alternative="greater")
```

As the confidence intervals are calculated on a logarithmic scale, taking the exponent of point- and confidence limit estimates yields in inference for the ratio of Poisson means.

```
> exp(ci$confint)
```

To compute the QA intervals, the function `glht` in package `multcomp` can be used, or the profile can be calculated by

```
> QA <- wald(mcp)
> ciQA <- confint(QA, alternative="greater")
> exp(ciQA$confint)
```

Likewise the HOA intervals are obtained:

```
> HOA <- hoa(mcp)
```

```
> ciHOA <- confint(HOA, alternative="greater")
> exp(ciHOA$confint)
```

A graphical comparison of all three profiles is available by the function

```
> compareplot(mcp, QA, HOA)
```

plotting all profile curves together in one graphic.

Table 5.3: Lower confidence limits for many-to-one comparisons for the micronucleus data example.

		Lower limits		
Comparison	Estimate	QA	SRDP	HOA
Hydro_{30} / Vehicle	1.48	0.76	0.75	0.76
Hydro_{50} / Vehicle	2.41	1.32	1.34	1.33
Hydro_{75} / Vehicle	5.44	3.18	3.26	3.25
Hydro_{100} / Vehicle	7.78	4.63	4.76	4.73

As ratios of Poisson means are observed, lower confidence limits larger than 1 denote significant differences to the control, which is first occurring for dose Hydro_{50} (Table 5.3). The exponent of the estimates for the Williams-type comparisons (Table 5.4), using the contrast

```
> K <- contrMat(table(Mutagenicity$Treatment), type="Williams")
```

are not this easy to interpret, as constructing differences of pooled groups of parameters on the logarithmic scale yields ratios of geometric means on the transformed scale. Nevertheless, a significant trend can be detected.

Table 5.4: Lower confidence limits for Williams-type trend detection for the micronucleus data example.

		Lower limits		
Comparison	Estimate	QA	SRDP	HOA
C 1	7.78	4.89	5.01	4.97
C 2	6.51	4.15	4.26	4.23
C 3	4.67	2.98	3.05	3.03
C 4	3.50	2.23	2.28	2.27

No large differences between the deviance profiles and the quadratic approximation can be seen for this relatively large number of counts. This is an assurance, that the quadratic approximation can safely be used for this dataset, with probably good performance in terms of coverage probability.

5.2.3 Cell transformation assay

Likewise to the micronucleus assay data (Section 2.3), a quasipoisson model is assumed for the counts in the cell transformation assay. The extra variability due to the sampling from several dishes is included by a single overdispersion parameter. This parameter is estimated as $\hat{\phi} = 1.6$, thus a small amount of extra variability has been found.

The R code is quite similar to the micronucleus assay data:

```
> str(dat)
'data.frame': 80 obs. of  2 variables:
 $ trt : Factor w/ 8 levels "solvent","0.01",..: 1 1 1 1 1 1 ...
 $ foci: int  0 1 0 0 0 0 0 0 0 2 ...
> fit <- glm(foci ~ trt-1, data=dat, family=quasipoisson(link="log"))
> K <- contrMat(table(dat$trt), type="Dunnett")
> #K <- contrMat(table(dat$trt), type="Williams")
> mcp <- mcprofile(fit, K)
> ci <- confint(mcp, alternative="greater")
> exp(ci$confint)

> QA <- wald(mcp)
> ciQA <- confint(QA, alternative="greater")
> exp(ciQA$confint)

> HOA <- hoa(mcp)
> ciHOA <- confint(HOA, alternative="greater")
> exp(ciHOA$confint)
```

Table 5.5: Lower confidence limits for many-to-one comparisons for the first cell transformation assay example.

		Lower limits		
Comparison	Estimate	QA	SRDP	HOA
0.01 / solvent	1.67	0.23	0.23	0.23
0.03 / solvent	1.00	0.11	0.09	0.09
0.1 / solvent	4.00	0.70	0.85	0.82
0.3 / solvent	8.67	1.67	2.18	2.11
1 / solvent	15.33	3.07	4.10	4.00
3 / solvent	16.67	3.35	4.48	4.37
10 / solvent	19.67	3.98	5.35	5.22

The lowest dose with a significant difference to the control can be found at dose 0.3 for all approaches (Table 5.5). Additionally a significant trend can be found by Williams-type trend tests shown in Table 5.6. As the spontaneous rate in the control is relatively low and only two Type 3 foci are counted, a quadratic approximation to the deviance profile might

loose some information about the profile curvature. Thus, different confidence limits are found, especially far in the alternative of a difference to the control or a highly significant trend. Here, the profile methods provide more clear decisions, showing a higher power.

Table 5.6: Lower confidence limits for Williams-type trend detection for the first cell transformation assay example.

		Lower limits		
Comparison	Estimate	QA	SRDP	HOA
C 1	19.67	5.09	6.33	6.19
C 2	18.10	4.75	5.96	5.84
C 3	17.13	4.52	5.69	5.58
C 4	14.45	3.82	4.80	4.71
C 5	11.17	2.94	3.70	3.63
C 6	7.47	1.94	2.42	2.36
C 7	6.03	1.57	1.95	1.91

The second dataset is an extreme example with no foci counted at all in the control. Estimating the mean for this group on the log link requires a stopping rule at some low parameter value to reach rough convergence. The dispersion parameter is estimated at $\hat{\phi} = 0.87$. This estimated variability, lower than under a Poisson assumption, is hard to justify in this context. Moreover, the estimation of the dispersion parameter might been unstable due to the zero control group. Therefore, a simple Poisson model is used with the dispersion parameter set to $\phi = 1$.

The R code is not much different than for the first cell transformation dataset. The profile is computed automatically for every parameter in a range from -20 to 20 as the control group consists of only zero counts.

```
> str(dat)
'data.frame': 80 obs. of  6 variables:
 $ trt  : Factor w/ 8 levels "Solvent","0.01","0.03",..: 1 1 1 1 1 1 ...
 $ foci : int  0 0 0 0 0 0 0 0 0 0 ...
> fit <- glm(foci ~ trt-1, data=dat, family=poisson(link="log"))
> K <- contrMat(table(dat$trt), type="Dunnett")
> mcp <- mcprofile(fit, K)
...
```

Instead of searching for minimal and maximal values automatically to restrict the profile calculations, they can directly be specified by a limits argument and distributing the supporting points equally over this range.

```
> lims <- cbind(rep(-5,nrow(K)), rep(10,nrow(K)))
```

```
> mcp <- mcprofile(fit, K, limits=lims, equally.spaced=TRUE)
...
```

Table 5.7: Lower confidence limits for many-to-one comparisons for the second cell transformation assay example.

		Lower limits		
Comparison	Estimate	QA	SRDP	HOA
0.01 / solvent	17772221	0	1.03	0
0.03 / solvent	26658332	0	1.76	0
0.1 / solvent	88861105	0	6.90	0
0.3 / solvent	79974995	0	6.17	0
1 / solvent	346558310	0	28.33	0
3 / solvent	790863836	0	65.29	0
10 / solvent	1874969320	0	155.47	0

The comparisons to the control (Table 5.7) and the Williams-type comparisons (Table 5.8) generate enormously high parameter estimates. The outcome are large variance estimates yielding in lower confidence limits of zero for the QA and HOA method. Different results can be obtained from the SRDP, showing a significant difference already at the comparison of the lowest dose to the control and a significant trend. These confidence intervals might be liberal as seen in the simulation study, but at least some useful limits could be obtained.

Table 5.8: Lower confidence limits for Williams-type trend detection for the second cell transformation assay example.

		Lower limits		
Comparison	Estimate	QA	SRDP	HOA
C 1	1874969320	0	155.47	0
C 2	1217721409	0	101.03	0
C 3	800984416	0	66.42	0
C 4	450252990	0	37.21	0
C 5	325467499	0	26.89	0
C 6	214484139	0	17.65	0
C 7	150270336	0	12.31	0

5.2.4 Liatrozole dose response study

The liatrozole study, introduced in Section 2.4, has a simple 4×2 table as outcome. The probability of observing an improvement at each dose is estimated in a binomial GLM on the logit link.

```
> str(dat)
```

```
'data.frame': 4 obs. of  3 variables:
 $ Dose      : Factor w/ 4 levels "Placebo","50",..: 1 2 3 4
 $ Improved  : num  2 6 4 13
 $ Unaffected: num  32 27 32 21
> fit <- glm(cbind(Improved, Unaffected) ~ Dose-1, data=dat,
+                          family=binomial(link="logit"))
> K <- contrMat(table(dat$Dose), type="Dunnett")
> mcp <- mcprofile(fit, K)
> ci <- confint(mcp, alternative="greater")
> exp(ci$confint)
```

Table 5.9: Lower confidence limits for many-to-one comparisons for the liatrozole example.

		Lower limits		
Comparison	Estimate	QA	SRDP	HOA
OR: 50 vs. Placebo	3.56	0.65	0.73	0.70
OR: 75 vs. Placebo	2.00	0.33	0.35	0.34
OR: 100 vs. Placebo	9.90	1.98	2.38	2.28

Applying a Dunnett contrast and exponentiating the results, yields in lower confidence intervals for the odds ratio for dose groups versus the control. The lower limits presented in Table 5.9 indicate a significant difference only for the 100mg dose to the placebo group. A significant Williams-trend can also be detected (Table 5.10) on the logit transformed scale. The estimates for the Williams contrasts are some kind of 'pooled odds-ratios', constructing linear combinations of parameter estimates on the logit scale and then taking the exponent.

Table 5.10: Lower confidence limits for Williams-type trend detection for the liatrozole example.

		Lower limits		
Comparison	Estimate	QA	SRDP	HOA
OR: C 1	9.90	2.24	2.64	2.52
OR: C 2	4.45	1.04	1.22	1.16
OR: C 3	4.13	1.00	1.20	1.14

Due to the small sample sizes, large differences between the SRDP and the quadratic approximation can be found. For the third Williams contrast the QA limits are only on the border of showing a significant trend; the SRDP and HOA limits are noticeably located in the alternative.

5.2.5 Fetal deaths after exposure to boric acid

The last example, introduced in Section 2.5, has a more complex layout, which can be summarized by an 4×2 table, but with additional replicates for each dose group. To capture the animal variability, a GLM under quasibinomial assumption can be used to estimate the proportions of dead fetuses per dose on the logit link. The boxplot in Figure 2.5 indicates a higher variability for the animals at the highest dose. Summarizing the animal variability by a single dispersion parameter over all doses might be misleading, and estimating several dispersion parameters by extended quasi-likelihood can be a better choice. But then the accuracy of the estimation might suffer from a decreased residual degree of freedom. The single dispersion parameter is estimated at $\hat{\phi} = 1.93$ showing a fair amount of animal variability.

```
> str(dat)
'data.frame': 107 obs. of  4 variables:
 $ Dose : Factor w/ 4 levels "control","low",..: 1 1 1 1 1 1 ...
 $ dead : num  0 0 1 1 1 2 0 0 1 2 ...
 $ alive: num  15 3 8 11 12 11 16 11 10 6 ...
> fit <- glm(cbind(Improved, Unaffected) ~ Dose-1, data=dat,
+                          family=quasibinomial(link="logit"))
> K <- contrMat(table(dat$Dose), type="Dunnett")
> mcp <- mcprofile(fit, K)
...
```

Table 5.11: Lower confidence limits for many-to-one comparisons for the fetal deaths example.

		Lower limits		
Comparison	Estimate	QA	SRDP	HOA
OR: low vs. control	1.28	0.56	0.56	0.56
OR: middle vs. control	0.74	0.29	0.28	0.28
OR: high vs. control	3.46	1.66	1.71	1.70

By the many-to-one comparisons in Table 5.11 the high dose can be classified being significantly different to the control. Also a significant trend can be detected by the Williams contrasts (Table 5.12).

Table 5.12: Lower confidence limits for Williams-type trend detection for the fetal deaths example.

		Lower limits		
Comparison	Estimate	QA	SRDP	HOA
OR: C 1	3.46	1.76	1.80	1.80
OR: C 2	1.58	0.80	0.81	0.81
OR: C 3	1.47	0.77	0.79	0.79

No large differences between the QA and the profile limits can be detected, assuring the safe use of the Wald intervals.

Chapter 6

Discussion

Deviance profiles are a useful method to visualize the uncertainty about a parameter estimate in a compact way, treating all others as nuisance parameters. The comparison of the profile with an adequate quadratic function is an indicator if the standard Wald methods for inference yield appropriate results. Furthermore, the comparison of profiles for two different parameters provide a way to roughly identify the correlation between them. But profiling cannot only be used to arrange nice displays; the direct relationship to a likelihood-ratio test makes them a useful tool for parametric statistical inference by tests and confidence intervals. Not only tests of global hypothesis about model parameters are available; signed root deviance profile based confidence intervals enable the application of profiling also to multiple contrast methods with adequate control of an approximate global type-I-error rate. The transformation invariance is a useful feature of the likelihood based methods, allowing for inference for functions of parameters without loss of information. Confidence intervals based on a quadratic approximation the the likelihood will produce different results, depending on the link function in a GLM. For example, the calculation of confidence intervals for the difference of proportions in a Binomial GLM on the identity link, a risk ratio on the log link, and an odds-ratio on the logit link, might result in different conclusions for the Wald method. The profile deviance methods will yield the same results independent of the chosen scale of the parameters.

The simulation results show the applicability of the methods to a broad range of settings. The coverage probability of the profile based simultaneous confidence intervals is maintained more closely at the designated level compared to common Wald intervals.

Especially, at small sample sizes improved confidence intervals are obtained. But overall a more liberal performance of the SRDP intervals can be determined; for problems that require more conservative decisions one may prefer the Wald method or, as a compromise, the higher order approximations.

The largest difference between the methods can be found for samples, where no event is counted at all. The SRDP based intervals use all available information of the observed data to provide an estimate of a confidence limit for the difference of sample means. This method tends to underestimate the variance for these parameters of interests, leading to liberal confidence intervals, as it is seen in the simulation study. Also the approximation with a normal distribution might be inappropriate. At least some estimates for a confidence limit can be obtained for these settings; if a more conservative approach is needed, the corresponding limits may be set to $[-\infty; \infty]$.

A major drawback of the profiling methods is their high computational effort. Simple models with a small number of parameters can easily be handled; for more complex problems it may not be worth to spend the additional costs of constructing a profile.

There are several alternatives to conditional likelihood inference available. An even more flexible but also more time consuming approach are resampling or bootstrap methods, e.g. introduced in Efron and Tibshirani (1994). As resampling methods are also directly based on the observed data, inference for a parameter is only available conditional on the specific sample. Most appealing is the applicability of resampling methods to almost any problem; also a combination of resampling and profiling methods are possible (Brazzale et al., 2007).

Instead of following the frequentist paradigm of believing in a true parameter and constructing confidence intervals that contain this parameter in $100(1-\alpha)\%$ of repeated sampling from an entire population, Bayesian intervals can be constructed, e.g. described in Gelman et al. (2003). Assuming a model with prior knowledge about the distribution of the parameters, the information of a random sample of observations can be used to support these assumptions. By weighting the prior distribution with the likelihood obtained from the sample, a posterior distribution of the parameter of interest is obtained. Confidence intervals based on this posterior distribution will contain the parameter of interest with probability $1-\alpha$, which might match the experimental question more pre-

cisely. Using uninformative prior distributions will lead to confidence intervals with quite similar frequentist coverage probability as the profile methods, because only the likelihood of the observed data is used. Especially, if the information provided by the observed data is low, providing an adequate prior distribution may improve the results a lot.

As an outlook, there might be several extensions of the profile methods possible, like applying them on (generalized) nonlinear models, as here a large improvement might be expected towards the standard Wald-type methods. Also the construction of profiles for ratios of model parameters, like it is done as an extension to Fieller's method in Dilba et al. (2006), may be of interest for future work.

Bibliography

Adler, I., Kliesch, U., 1990. Comparison of single and multiple treatment regimens in the mouse bone-marrow micronucleus assay for hydrochinone (HQ) and cyclophosphamide (CP). *Muatation Research* **234** (3-4), 115–123.

Agresti, A., 2007. *An Introduction to Categorical Data Analysis* (Wiley Series in Probability and Statistics), 2nd Edition. Wiley-Interscience.

Agresti, A., Bini, M., Bertaccini, B., Ryu, E., 2008. Simultaneous confidence intervals for comparing binomial parameters. *Biometrics* 64 (4), 1270–1275.

Bailer, A. J., Piegorsch, W., 1997. *Statistics for Environmental Biology and Toxicology* (Interdisciplinary Statistics), 1st Edition. Chapman & Hall/CRC.

Barndorff-Nielsen, O., 1986. Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73** (2), 307–322.

Barndorff-Nielsen, O., Cox, D., 1979. Edgeworth and saddle-point approximations with statistical applications. *Journal of the Royal Statistical Society Series B-Methodological* **41** (3), 279–312.

Bates, D. M., Watts, D. G., 1988. *Nonlinear Regression Analysis and Its Applications*. Wiley.

Berth-Jones, J., Todd, G., Hutchinson, P., Thestrup-Pedersen, K., Vanhoutte, F., 2000. Treatment of psoriasis with oral liarozole: a dose-ranging study. *British Journal of Dermatology* **143** (6), 1170–1176.

Boyles, R. A., 2008. The role of likelihood in interval estimation. *American Statistician* **62** (1), 22–26.

Brazzale, A. R., Davison, A. C., 2008. Accurate Parametric Inference for Small Samples. *Statistical Science* **23** (4), 465–484.

Brazzale, A. R., Davison, A. C., Reid, N., 2007. *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge University Press.

Bretz, F., 2006. An extension of the Williams trend test to general unbalanced linear models. *Computational Statistics & Data Analysis* **50** (7), 1735–1748.

Bretz, F., Genz, A., Hothorn, L., 2001. On the numerical availability of multiple comparison procedures. *Biometrical Journal* **43** (5), 645–656.

Chen, J., Jennrich, R., 1996. The signed root deviance profile and confidence intervals in maximum likelihood analysis. *Journal of the American Statistical Association* **91** (435), 993–998.

Chen, J., Jennrich, R., 2002. Simple accurate approximation of likelihood profiles. *Journal of Computational and Graphical Statistics* **11** (3), 714–732.

Czado, C., Munk, A., 2000. Noncanonical links in generalized linear models - when is the effort justified? *Journal of Statistical Planning and Inference* **87** (2), 317–345.

Davison, A., 1988. Approximate conditional inference in generalized linear-models. *Journal of the Royal Statistical Society Series B-Methodology* **50** (3), 445–461.

Dilba, G., Bretz, F., Guiard, V., 2006. Simultaneous confidence sets and confidence intervals for multiple ratios. *Journal of Statistical Planning and Inference* 136 (8), 2640–2658.

Donaldson, J., Schnabel, R., 1987. Computational experience with confidence-regions and confidence intervals for nonlinear least-squares. *Technometrics* **29** (1), 67–82.

Dunnett, C., 1955. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50** (**272**), 1096–1121.

Efron, B., Tibshirani, R., 1994. *An Introduction to the Bootstrap* (Monographs on Statistics and Applied Probability), 1st Edition. Chapman & Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 2003. *Bayesian Data Analysis* (Texts in Statistical Science), 2nd Edition. Chapman & Hall/CRC.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T., 2009. mvtnorm: Multivariate normal and t distributions. R package version 0.9-7.
URL `http://CRAN.R-project.org/package=mvtnorm`

Hauck, W., Donner, A., 1977. Walds test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association* 72 (360), 851–853.

Hirotsu, C., Marumo, K., 2002. Changepoint analysis as a method for isotonic inference. *Scandinavian Journal of Statistics* **29** (1), 125–138.

Hochberg, Y., Tamhane, A., 1987. *Multiple comparison procedures*. John Wiley & Sons, Inc.

Hothorn, T., Bretz, F., Westfall, P., 2008. Simultaneous inference in general parametric models. *Biometrical Journal* **50** (3), 346–363.

Johnson, N. L., Kotz, S., Balakrishnan, N., 1994. *Continuous Univariate Distributions, Vol. 1* (Wiley Series in Probability and Statistics), 2nd Edition. Wiley-Interscience.

Johnson, N. L., Kotz, S., Balakrishnan, N., 1995. *Continuous Univariate Distributions, Vol. 2* (Wiley Series in Probability and Statistics), 2nd Edition. Wiley-Interscience.

Lawless, J., 1987. Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics-Revue Canadienne De Statistique* **15** (3), 209–225.

Lindsey, J. K., 1996. *Parametric Statistical Inference*. Oxford University Press, USA.

McCullagh, P., Nelder, J. A., 1989. *Generalized Linear Models*, (Monographs on Statistics and Applied Probability), 2nd Edition. Chapman & Hall/CRC.

Nelder, J., Wedderburn, R., 1972. Generalized linear models. *Journal of the Royal Statistical Society Series A-General* **135** (3), 370–&.

Piegorsch, W., 1991. Multiple comparisons for analyzing dichotomous response. *Biometrics* 47 (1), 45–52.

Pierce, D., Peters, D., 1992. Practical use of higher-order asymptotics for multiparameter exponetial-families. *Journal of the Royal Statistical Society Series B-Methodology* **54** (3), 701–737.

Quinn, S., Bacon, D., Harris, T., 1999. Assessing the precision of model predictions and

other functions of model parameters. *Canadian Journal of Chemical Engineering* **77** (4), 723–737.

Quinn, S., Bacon, D., Harris, T., 2000. Notes on likelihood intervals and profiling. *Communications in Statistics - Theory and Methods* **29** (1), 109–129.

R Development Core Team, 2009. R: A language and environment for statistical computing. ISBN 3-900051-07-0.
URL `http://www.R-project.org`

Reid, N., 2000. Likelihood. *Journal of the American Statistical Association* **95** (452), 1335–1340.

Schaarschmidt, F., Sill, M., Hothorn, L. A., 2008. Approximate Simultaneous Confidence Intervals for Multiple Contrasts of Binomial Proportions. *Biometrical Journal* 50 (5), 782–792.

Skovgaard, I., 2001. Likelihood asymptotics. *Scandinavian Journal of Statistics* **28** (1), 3–32, 17th Nordic Conference on Mathematical Statistics, Helsingor, Denmark, JUN, 1998.

Skovgaard, I. M., 1996. An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2** (2), 145–165.

Staicu, A.-M., 2009. Higher-order approximations for interval estimation in binomial settings. *Journal of Statistical Planning and Inference* **139** (10), 3393–3404.

Thomas, C., 2009. Unpublished cell transformation assay data. Provided in course of the ECVAM JRC project 'Quality assessment and novel statistical analysis techniques for toxicological data'. CCR.IHCP.C433223.X0.

Tukey, J. W., 1983. *The problem of multiple comparisons*, in Collected Works of John W. Tukey: VIII Multiple Comparisons, 1948-1983.

Venables, W., Ripley, B., 2003. *Modern Applied Statistics with S*, 4th Edition. Springer.

Venzon, D., Moolgavkar, S., 1988. A method for computing profile-likelihood-based confidence-intervals. *Applied Statistics - Journal of the Royal Statistical Society Series C* **37** (1), 87–94.

Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. *Annals of mathematical statistics* **20** (4), 595–601.

Wedderburn, R., 1974. Quasi-likelihood functions, generalized linear-models, and Gauss-Newton method. *Biometrika* **61** (3), 439–447.

Westfall Peter H., Tobias Randall D., R. D. W. R. D. H. Y., 1999. *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Publishg.

Appendix A

Appendix

A.1 Commonly used deviance functions

Gaussian deviance:

$$D(\mu_i, y_i) = 2 \sum_{i=1}^{N} (y_i - \mu_i)^2$$

Poisson deviance:

$$D(\mu_i, y_i) = 2 \sum_{i=1}^{N} w_i \left\{ y_i \log \left(\gamma - (y_i - \mu_i) \right) \right\},$$

with a weight vector w_i and $\gamma = \begin{cases} \frac{y_i}{\mu_i} & \text{if } y \neq 0 \\ 1 & \text{y} = 0 \end{cases}$

Binomial deviance:

$$D(\mu_i, y_i) = 2 \sum_{i=1}^{N} w_i \left\{ y_i \log (\gamma) + (1 - y_i) \log (\kappa) \right\},$$

with a weight vector w_i, $\gamma = \begin{cases} \frac{y_i}{\mu_i} & \text{if } y > 0 \\ 1 & \text{y} = 0 \end{cases}$, and $\kappa = \begin{cases} \frac{1-y_i}{1-\mu_i} & \text{if } y < 1 \\ 1 & \text{y} = 1 \end{cases}$

Negative binomial deviance:

$$D(\mu_i, y_i) = 2 \sum_{i=1}^{N} \left\{ y_i \log (\gamma) - (y_i + \rho) \log \left(\frac{y_i + \rho}{\mu_i + \rho} \right) \right\},$$

with a dispersion parameter ρ and $\gamma = \begin{cases} \frac{y_i}{\mu_i} & \text{if } y \geq 1 \\ \frac{1}{\mu_i} & \text{y} < 1 \end{cases}$

A.2 Exemplary contrast matrices

Dunnett:

$$c_{mj} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Tukey:

$$c_{mj} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

Williams:

$$c_{mj} = \begin{bmatrix} -1 & 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 1/2 & 1/2 \\ -1 & 0 & 1/3 & 1/3 & 1/3 \\ -1 & 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

Changepoint:

$$c_{mj} = \begin{bmatrix} -1 & 1/4 & 1/4 & 1/4 & 1/4 \\ -1/2 & -1/2 & 1/3 & 1/3 & 1/3 \\ -1/3 & -1/3 & -1/3 & 1/2 & 1/2 \\ -1/4 & -1/4 & -1/4 & -1/4 & 1 \end{bmatrix}$$

A.3 Additional simulation results

A.3.1 Count data

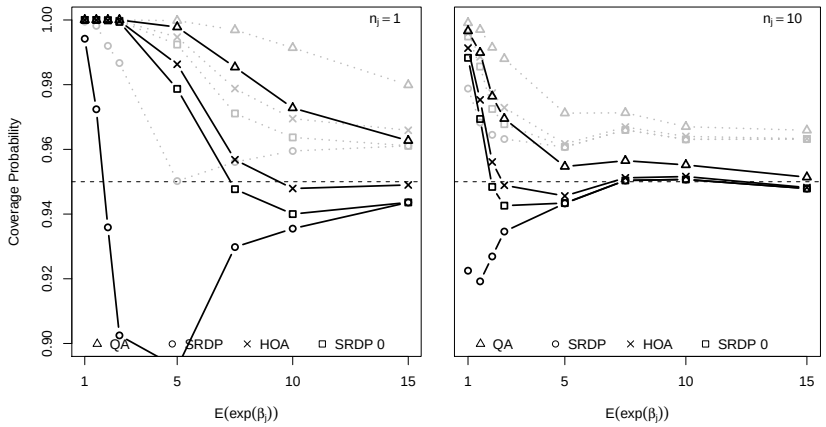

Figure A.1: One-sided Dunnett-type comparisons for 5 ordered means estimated in a Poisson GLM. Simulation settings are varied by the mean of a Gamma distribution $E(\beta_j)$ on the logarithmic scale and the number of replicates per group n_j. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *grey lines* represent multiplicity adjustment by Bonferroni.

A.3.2 Categorical data

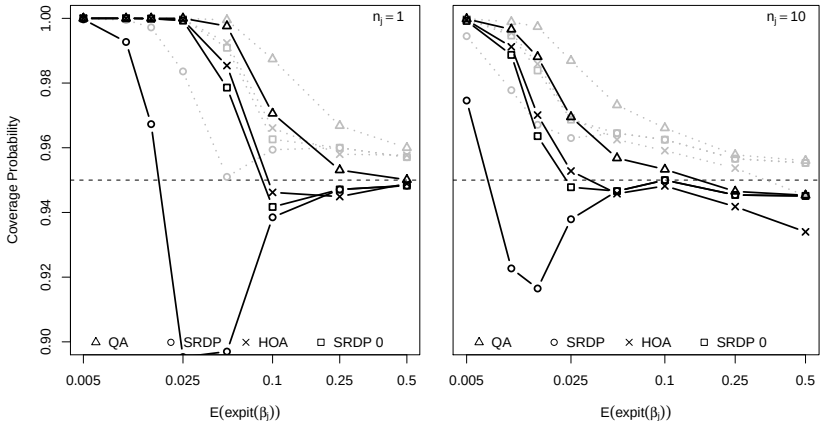

Figure A.2: One-sided Dunnett-type comparisons for 5 ordered parameters estimated in a Binomial GLM. Simulation settings are varied by the mean of a Beta distribution $E(\beta_j)$ on the logit link and the number of replicates per group n_j. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *grey lines* represent multiplicity adjustment by Bonferroni.

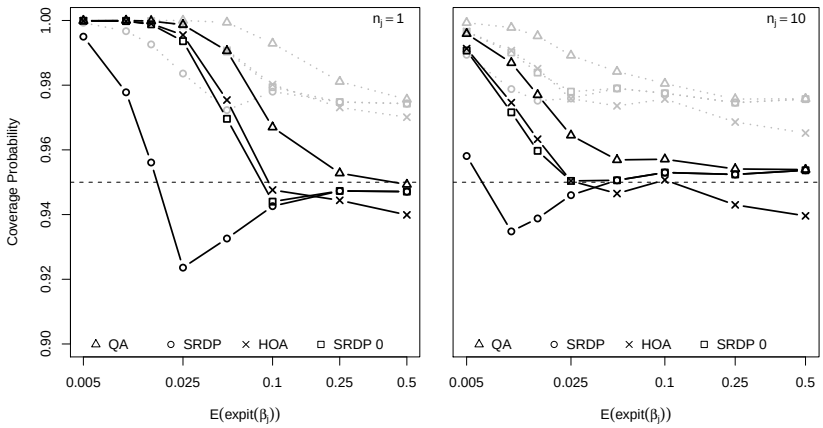

Figure A.3: One-sided Williams-type comparisons for 5 ordered parameters estimated in a Binomial GLM assuming a linear trend. Simulation settings are varied by the mean of a Beta distribution $E(\beta_j)$ on the logit link and the number of replicates per group n_j. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *gray lines* represent multiplicity adjustment by Bonferroni.

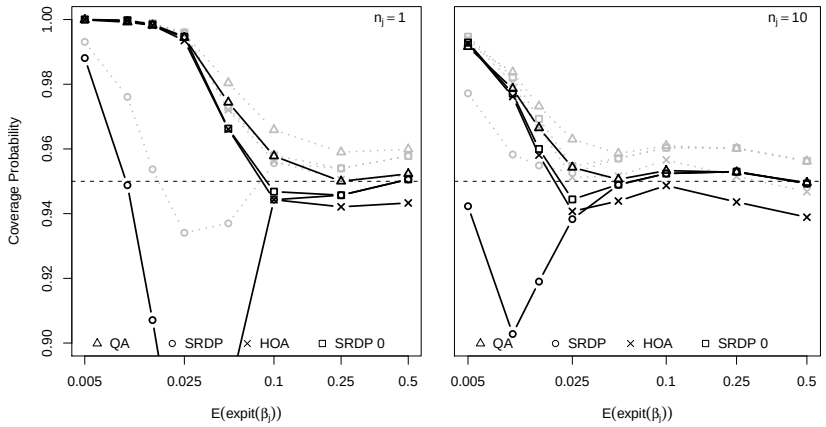

Figure A.4: Changepoint comparisons for 5 parameters estimated in a Binomial GLM. Simulation settings are varied by the mean of a Beta distribution $E(\beta_j)$ on the logit link and the number of replicates per group n_j. *Black lines* indicate multiplicity adjustment by a multivariate normal distribution with estimated correlation structure; *gray lines* represent multiplicity adjustment by Bonferroni.

A.4 R Code

A.4.1 Estimating SRDP

The following objects are needed:

x A vector of $k-1$ starting values (obtained from the original model)

X A design matrix

y A response vector

family A GLM family object

weight A vector of weights

deviance The deviance of the original model

dispersion The estimated dispersion parameter

k A vector with a single contrast

w Index of the restricted parameter

p Value of the parameter restriction

Then a constrained optimization can be performed, calling the function `optim` to optimize the function optfkt:

```
constDeviance <- function(x, X, y, family, weight){
  mu <- fam$linkinv(X %*% x)
  if (fam$validmu(fv)) sum(family$dev.resids(y=y, mu=mu, wt=weight))
        else 9999999999
```

```
}
srd <- function(x, X, y, deviance, dispersion, family, weight){
  sqrt(abs(constDeviance(x=x, X=X, y=y,
         fam=family, weight=weight)-deviance)/dispersion)
}
kbb <- function(x, k, p, w){
  K <- -1*c(-p, k[-w])/k[w]
  hv <- c(1,x)
  b1 <- (K %*% hv)[1,1]
  b <- numeric(length=length(k))
  b[(1:length(k))[-w]] <- x
  b[w] <- b1
  b
}
optfkt <- function(x){
  beta <- kbb(x=x, k=k, p=p, w=w)
  srd(beta, X=X, y=y, deviance=deviance,
          dispersion=dispersion, family=family, weight=weight)
}
```

A.4.2 Higher order approximation

Providing a `glm` and a `mcprofile` object with a list of estimated parameters `profile`, a profile based on higher order approximations is calculated by:

```
dKj <- diag((k == 0)[1,])
Kj <- rbind(k, dKj[apply(dKj,1,sum) != 0,])
vcfit <- summary(fit)$cov.unscaled
j <- 1/det(Kj %*% vcfit %*% t(Kj))
j.1 <- j / (1/(k %*% vcfit %*% t(k))[1,1])
ind <- k == 0
Xi <- X[,ind,drop=FALSE]
mp <- na.omit(as.data.frame(t(apply(profile,1,function(p){
  bp <- p[-c(1,2)]
  r <- -p[2]*sqrt(dispersion)
  d <- p[1]
  q <- ((k %*% b)[1,1] - dx) / sqrt(k %*% vcfit %*% t(k))[1,1]
  o <- X[,!ind,drop=FALSE] %*% unlist(bp[!ind])
  fm <- glm.fit(x = Xi, y = y, weights = weights,
         etastart = X %*% b, offset = o, family = family,
         control=fit$control)
  class(fm) <- "glm"
  csummary <- summary(fm)
  j.0 <- 1 / det(csummary$cov.unscaled)
  rho <- sqrt(j.1 / j.0)
  rstar <- (r + log(rho*q/r) / r)*1/sqrt(disp)
  rstar
}
```

A.4.3 Confidence intervals and tests

Quantiles of a multivariate t- or normal distribution are found with `vc` being the variance-covariance matrix of the estimated parameters and `K` being the specified contrast matrix.

```
require(mvtnorm)
VC <- K %*% vc %*% t(K)
d <- 1/sqrt(diag(VC))
dd <- diag(d)
cr <- dd %*% VC %*% dd
if (alternative == "two.sided"){
  if (is.null(df)){
    quantile <- qmvnorm(level, interval=c(-10, 10),
                corr=cr, tail="both.tails")$quantile
  } else {
    quantile <- qmvt(level, df=df, interval=c(-10, 10),
                corr=cr, tail="both.tails")$quantile
  }
} else {
  if (is.null(df)){
    quantile <- qmvnorm(level, interval=c(-10, 10),
                corr=cr, tail="lower.tail")$quantile
  } else {
    quantile <- qmvt(level, df=df, interval=c(-10, 10),
                corr=cr, tail="lower.tail")$quantile
  }
}
```

With `psi` being a vector of parameter values and `tau` the corresponding SRDP statistics, confidence intervals are found by interpolation with a spline function `sf()` :

```
sf <- splinefun(tau, psi))
mintau <- tau[1]
maxtau <- tau[length(tau)]
if ((alternative == "greater" | alternative == "two.sided") &
     mintau <= -quantile){
  lower <- sf(-quantile)
} else {
  lower <- -Inf
}
if ((alternative == "less" | alternative == "two.sided") &
     maxtau >= quantile){
  upper <- sf(quantile)
} else {
  upper <- Inf
}
```

Tests are constructed correspondingly for a margin `delta` by:

```
sf <- splinefun(psi, tau)
stat <- -sf(delta)
```

```
pfct <- function(q) {
  switch(alternative, two.sided = {
      low <- rep(-abs(q), dim)
      upp <- rep(abs(q), dim)
    }, less = {
      low <- rep(q, dim)
      upp <- rep(Inf, dim)
    }, greater = {
      low <- rep(-Inf, dim)
      upp <- rep(q, dim)
  })
  if (is.null(df)){
    pmvnorm(lower = low, upper = upp, corr = cr)
  } else {
    pmvt(lower = low, upper = upp, df=df, corr = cr)
  }
}
# for a vector of test statistics
padj <- numeric(length(stat))
for (i in 1:length(stat)) {
  padj[i] <- 1 - pfct(stat[i])
}
```

(This R Code is in some extend taken from the R package `multcomp` (Hothorn et al., 2008))