National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

## NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.



## AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.



Canada

# APPROXIMATIONS FOR MARGINAL DENSITIES OF $M$-ESTIMATORS

By

Rocky Yuk-Keung Fan

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
DALHOUSIE UNIVERSITY
HALIFAX, NOVA SCOTIA
AUGUST 23, 1994

ISBN 0-315-98842-8

Name ROCKY YUK-KEUNG FAN

*Dissertation Abstracts International* is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

STATISTICS — SUBJECT TERM

0463 — SUBJECT CODE

U·M·I

## Subject Categories

# THE HUMANITIES AND SOCIAL SCIENCES

### COMMUNICATIONS AND THE ARTS
- Architecture 0729
- Art History 0377
- Cinema 0900
- Dance 0378
- Fine Arts 0357
- Information Science 0723
- Journalism 0391
- Library Science 0399
- Mass Communications 0708
- Music 0413
- Speech Communication 0459
- Theater 0465

### EDUCATION
- General 0515
- Administration 0514
- Adult and Continuing 0516
- Agricultural 0517
- Art 0273
- Bilingual and Multicultural 0282
- Business 0688
- Community College 0275
- Curriculum and Instruction 0727
- Early Childhood 0518
- Elementary 0524
- Finance 0277
- Guidance and Counseling 0519
- Health 0680
- Higher 0745
- History of 0520
- Home Economics 0278
- Industrial 0521
- Language and Literature 0279
- Mathematics 0280
- Music 0522
- Philosophy of 0998
- Physical 0523
- Psychology 0525
- Reading 0535
- Religious 0527
- Sciences 0714
- Secondary 0533
- Social Sciences 0534
- Sociology of 0340
- Special 0529
- Teacher Training 0530
- Technology 0710
- Tests and Measurements 0288
- Vocational 0747

### LANGUAGE, LITERATURE AND LINGUISTICS
- Language
  - General 0679
  - Ancient 0289
  - Linguistics 0290
  - Modern 0291
- Literature
  - General 0401
  - Classical 0294
  - Comparative 0295
  - Medieval 0297
  - Modern 0298
  - African 0316
  - American 0591
  - Asian 0305
  - Canadian (English) 0352
  - Canadian (French) 0355
  - English 0593
  - Germanic 0311
  - Latin American 0312
  - Middle Eastern 0315
  - Romance 0313
  - Slavic and East European 0314

### PHILOSOPHY, RELIGION AND THEOLOGY
- Philosophy 0422
- Religion
  - General 0318
  - Biblical Studies 0321
  - Clergy 0319
  - History of 0320
  - Philosophy of 0322
- Theology 0469

### SOCIAL SCIENCES
- American Studies 0323
- Anthropology
  - Archaeology 0324
  - Cultural 0326
  - Physical 0327
- Business Administration
  - General 0310
  - Accounting 0272
  - Banking 0770
  - Management 0454
  - Marketing 0338
- Canadian Studies 0385
- Economics
  - General 0501
  - Agricultural 0503
  - Commerce Business 0505
  - Finance 0508
  - History 0509
  - Labor 0510
  - Theory 0511
- Folklore 0358
- Geography 0366
- Gerontology 0351
- History
  - General 0578
  - Ancient 0579
  - Medieval 0581
  - Modern 0582
  - Black 0328
  - African 0331
  - Asia, Australia and Oceania 0332
  - Canadian 0334
  - European 0335
  - Latin American 0336
  - Middle Eastern 0333
  - United States 0337
- History of Science 0585
- Law 0398
- Political Science
  - General 0615
  - International Law and Relations 0616
  - Public Administration 0617
- Recreation 0814
- Social Work 0452
- Sociology
  - General 0626
  - Criminology and Penology 0627
  - Demography 0938
  - Ethnic and Racial Studies 0631
  - Individual and Family Studies 0628
  - Industrial and Labor Relations 0629
  - Public and Social Welfare 0630
  - Social Structure and Development 0700
  - Theory and Methods 0344
- Transportation 0709
- Urban and Regional Planning 0999
- Women's Studies 0453

# THE SCIENCES AND ENGINEERING

### BIOLOGICAL SCIENCES
- Agriculture
  - General 0473
  - Agronomy 0285
  - Animal Culture and Nutrition 0475
  - Animal Pathology 0476
  - Food Science and Technology 0359
  - Forestry and Wildlife 0478
  - Plant Culture 0479
  - Plant Pathology 0480
  - Plant Physiology 0817
  - Range Management 0777
  - Wood Technology 0746
- Biology
  - General 0306
  - Anatomy 0287
  - Biostatistics 0308
  - Botany 0309
  - Cell 0379
  - Ecology 0329
  - Entomology 0353
  - Genetics 0369
  - Limnology 0793
  - Microbiology 0410
  - Molecular 0307
  - Neuroscience 0317
  - Oceanography 0416
  - Physiology 0433
  - Radiation 0821
  - Veterinary Science 0778
  - Zoology 0472
- Biophysics
  - General 0786
  - Medical 0760

### EARTH SCIENCES
- Biogeochemistry 0425
- Geochemistry 0996
- Geodesy 0370
- Geology 0372
- Geophysics 0373
- Hydrology 0388
- Mineralogy 0411
- Paleobotany 0345
- Paleoecology 0426
- Paleontology 0418
- Paleozoology 0985
- Palynology 0427
- Physical Geography 0368
- Physical Oceanography 0415

### HEALTH AND ENVIRONMENTAL SCIENCES
- Environmental Sciences 0768
- Health Sciences
  - General 0566
  - Audiology 0300
  - Chemotherapy 0992
  - Dentistry 0567
  - Education 0350
  - Hospital Management 0769
  - Human Development 0758
  - Immunology 0982
  - Medicine and Surgery 0564
  - Mental Health 0347
  - Nursing 0569
  - Nutrition 0570
  - Obstetrics and Gynecology 0380
  - Occupational Health and Therapy 0354
  - Ophthalmology 0381
  - Pathology 0571
  - Pharmacology 0419
  - Pharmacy 0572
  - Physical Therapy 0382
  - Public Health 0573
  - Radiology 0574
  - Recreation 0575
  - Speech Pathology 0460
  - Toxicology 0383
- Home Economics 0386

### PHYSICAL SCIENCES

#### Pure Sciences
- Chemistry
  - General 0485
  - Agricultural 0749
  - Analytical 0486
  - Biochemistry 0487
  - Inorganic 0488
  - Nuclear 0738
  - Organic 0490
  - Pharmaceutical 0491
  - Physical 0494
  - Polymer 0495
  - Radiation 0754
- Mathematics 0405
- Physics
  - General 0605
  - Acoustics 0986
  - Astronomy and Astrophysics 0606
  - Atmospheric Science 0608
  - Atomic 0748
  - Electronics and Electricity 0607
  - Elementary Particles and High Energy 0798
  - Fluid and Plasma 0759
  - Molecular 0609
  - Nuclear 0610
  - Optics 0752
  - Radiation 0756
  - Solid State 0611
- Statistics 0463

#### Applied Sciences
- Applied Mechanics 0346
- Computer Science 0984
- Engineering
  - General 0537
  - Aerospace 0538
  - Agricultural 0539
  - Automotive 0540
  - Biomedical 0541
  - Chemical 0542
  - Civil 0543
  - Electronics and Electrical 0544
  - Heat and Thermodynamics 0348
  - Hydraulic 0545
  - Industrial 0546
  - Marine 0547
  - Materials Science 0794
  - Mechanical 0548
  - Metallurgy 0743
  - Mining 0551
  - Nuclear 0552
  - Packaging 0549
  - Petroleum 0765
  - Sanitary and Municipal 0554
  - System Science 0790
- Geotechnology 0428
- Operations Research 0796
- Plastics Technology 0795
- Textile Technology 0994

### PSYCHOLOGY
- General 0621
- Behavioral 0384
- Clinical 0622
- Developmental 0620
- Experimental 0623
- Industrial 0624
- Personality 0625
- Physiological 0989
- Psychobiology 0349
- Psychometrics 0632
- Social 0451

*To My Parents*

# Contents

# Abstract

In this thesis we present a finite sample approximation for the marginal densities of a multivariate $M$-estimator. The result is particularly useful in robust statistics where an estimator usually is defined implicitly and does not have a closed form, and for small sample problems where the asymptotic results may not be reliable.

Precisely, let $Y_1, \ldots, Y_n$ be independent $m$-dimensional random observations such that each observation has a density function which is parameterized by a $p$-dimensional parameter $\eta$. Let $\hat{\eta}$ be an $M$-estimator of $\eta$, the solution of the system

$$\frac{1}{n}\sum_{l=1}^{n}\Psi_{jl}(Y_l, \hat{\eta}) = 0, \quad j = 1, \ldots, p.$$

Our primary objective is to derive an approximation for the marginal densities of a component in $\hat{\eta}$ under $\eta = \eta_0$. The result is then extended to a real-valued function $\rho(\hat{\eta})$, $\rho : \Re^p \rightarrow \Re$, and finally to a real-valued vector $\rho(\hat{\eta}) = \{\rho_1(\hat{\eta}), \ldots, \rho_k(\hat{\eta})\}$, $\rho : \Re^p \rightarrow \Re^k$, $k \leq p$.

We begin with an overview of the general problem and some background information. Then we derive the main results and discuss the relationship among our approach and some existing techniques for the problem. In addition, we implement the approximations for several location-scale and multiple regression examples. Finally, we discuss the limitation and some potential applications of our results.

# Acknowledgements

I would like to express my deepest gratitude and thanks to my supervisor, Dr. Christopher A. Field, for his guidance and patience over the past four years. His excellent supervision was essential to the successful completion of this thesis and will be invaluable throughout my academic career.

I would like to thank my external examiner, Dr. D.F. Andrews, for his valuable comments, and Dr. G. Gabor and Dr. L. Manchester for reading this thesis and providing very helpful suggestions. Their input has greatly improved the quality of the work.

I am grateful to the Faculty of Graduate Studies for awarding me Dalhousie Graduate Scholarships. My sincere thanks go to all those who have helped me in various stages of my studies. I particularly thank Miss Xiaowei Li for her help with my early exploration to the area. Thanks also go to the Graduate Coordinators, Dr. P.N. Stewart, Dr. K.P. Johnson and Dr. W.R.S. Sutherland, and to the secretarial team of the department.

# Chapter 1

# Introduction

## 1.1 Overview

The objective of this thesis is to develop an approximation for the marginal density function of a multivariate $M$-estimator. The result is particularly useful in robust statistics and for small sample applications.

Generally speaking, an estimator is a function defined by a set of random observations, which can be used to reveal a certain characteristic of a population. To use it in practice, we must know its random behaviour. Also, the knowledge is needed if we want to compare different classes of estimators. This leads us to think of a common source of the information, the distribution function.

When an explicit distribution function is available, we can use it to obtain the information that we need, otherwise, we have to compute it numerically. However, except for some simple functions of the random observations, computation of the exact distribution could be intractable. In fact, an estimator may only be defined implicitly and does not have a closed form. Unfortunately, most of the robust estimators are in this last category. Therefore, an approximation, or precisely, an accurate

approximation for the distribution is clearly needed.

In the classical theory, the computation of a distribution function may not be difficult. Under the assumption of normality, very nice and complete results under different settings have been found. Most of the statistics used in practice have very well known distributions, and even if they do not, asymptotic results are usually available to provide satisfactory alternatives.

However, the situation has been changing as robustness of classical results has become a concern to statisticians. We now realize that the arithmetic mean of a random sample is highly non-robust in the sense that a single outlier can cause the estimate to break down. As a result, different robust procedures have been developed in the last few decades. In particular, a general class of robust estimators was proposed by Huber (1964). The estimators are known as the $M$-estimators. In brief, an $M$-estimator is defined implicitly as the solution of a system of equations. Huber showed in the same paper that the new class of estimators possesses very desirable quantitative and qualitative properties.

Since they were introduced, the $M$-estimators have been the basis of new developments in robust procedures. Various modifications and extensions have been proposed, and their sampling behaviours, mostly in the asymptotic sense, have been explored. Since robust estimators usually cannot be computed analytically, it is difficult to study their finite sample properties, and therefore statistical inferences have to rely on asymptotic results.

Although asymptotic results usually are available in cases of interest, they may be inadequate for practical purposes. For instance, when an estimator is asymptotically normally distributed, we could use this result to approximate the true distribution.

Although the asymptotic normality is a very nice feature for an estimator, it does not always produce reasonable approximations unless the sample size is large enough. Even worse, we do not know how large is large enough in an individual case. We observe from numerical examples that a moderate size is possibly too risky. Nonetheless, the normal approximation tends to be quite reasonable around the center but it can be very inaccurate in the tails of the distribution.

Different techniques have been developed to provide more accurate approximations. In general, one can try to improve the asymptotic results or to approximate directly the finite sample distributions. In the latter case, there are options such as approximating the estimator itself or the distribution of the estimator. In particular, Field (1982) had successfully derived a very accurate approximation for the joint density function of a multivariate $M$-estimator. An important step in his approach is the use of the saddlepoint technique.

In a pioneering paper, Daniels in 1954 applied the saddlepoint technique and derived a very accurate approximation for the distribution of an arithmetic mean. In the last forty years, the technique has been proven to be very useful in small sample asymptotics, a name coined by Hampel. The name reflects the aim of obtaining asymptotic expansions which give accurate approximations for small samples.

For those problems where the marginal densities of a component in a multivariate estimator are needed, one may use the results in Field (1982) to approximate the joint densities of the components, and then integrate out the nuisance variables. This approach was demonstrated by Field in the same paper for a two-dimensional problem and gave very good results. However, the process involves substantial computational effort and becomes impractical when the dimension exceeds two.

Tingley and Field (1990) manipulated results in Field (1982) to derive a linear approximation to a real-valued function of the components. The problem is then reduced to one-dimension and the computation becomes feasible. The approximation was used as the basis for constructing robust confidence intervals in Tingley and Field (1990) and Tingley (1992). In spite of its simplicity, we see later that this single linear approximation may not provide satisfactory approximations in the tail regions of most interest.

In the present work, we develop an approximation which is reliable even well out into the tails. The work is motivated partly by the good performance of the linear approximation around the expected value of the estimator. Our approximation is applied to several robust multiple regression problems and yields very good results for small samples. This enables us to study the finite sample behaviour of an estimator and to compare it with the asymptotic results. From that, we can determine when asymptotic approximations are sensible for use in practice, or compare different estimators on the basis of their finite sample properties.

The remainder of this chapter contains background information on our work. The next section gives a brief review of $M$-estimators. The exponential tilt plays a central role in our approximation and is discussed in Section 1.3 Section 1.4 presents a normal approximation which is needed for local approximations in the development. Finally in the last section, we summarize the discussion and outline the content of the subsequent chapters.

## 1.2 $M$-estimator

The $M$-estimator is undoubtedly one of the more influential ideas in statistics within the last few decades. Numerous robust procedures have been inspired by it and developed based on it. In this section, we state the definition and two asymptotic properties of the estimator. The theoretical details are skipped in the discussion and can be found in Huber (1981).

In his paper in 1964, Huber introduced a general class of estimators which he called $M$-estimators. The proposed estimator was defined first as the solution of a minimization problem and then extended to more general situations. We now set some notation for this section and give the definition of the estimator.

Let $Y_1, \ldots, Y_n$ be a sample of size $n$, where each of the $Y_l$'s has a distribution function $F_l(y_l)$ parameterized by $\eta = (\eta_1, \ldots, \eta_p)$. The true value of $\eta$ is $\eta_0$. The density function, when it exists, is denoted by $f_l(y_l)$.

**Definition 1.1** (Huber, 1981, page 43)

*An $M$-estimator of $\eta \in \Omega \subset \Re^p$ is defined as the solution $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_p)$ of the minimization problem*

$$\sum_{l=1}^{n} \tau_l(Y_l, \hat{\eta}) = \min! \tag{1.1}$$

*over the parameter space $\Omega$, or implicitly as the solution of the system of $p$ equations*

$$\sum_{l=1}^{n} \Psi_l(Y_l, \hat{\eta}) = 0, \tag{1.2}$$

*where*

$$\Psi_l(Y_l, \eta) = \frac{\partial \tau_l(Y_l, \eta)}{\partial \eta} = \{\Psi_{1l}(Y_l, \eta), \ldots, \Psi_{pl}(Y_l, \eta)\}, \quad l = 1, \ldots, n,$$

*are called the score functions.*

□

Note that it is not assumed in the definition, but we require that the random observations to be independent in our development.

It was pointed out by Huber himself that the functional version of (1.1) may cause problems. For instance, the median $T$ of a random sample from a common distribution $F$ corresponds to

$$\tau_l(Y_l, T) = |Y_l - T|, \quad l = 1, \ldots, n,$$

but we cannot define it to be an estimator of $t$ that minimizes

$$\int_{y_l} |y_l - t| \, dF(y_l)$$

unless $Y_l$ has a finite first absolute moment. On the other hand, the definition in (1.2) may lead to multiple solutions which correspond to local minima of (1.1). For the above example, the score functions are

$$\Psi_l(Y_l, T) = -sign\{Y_l - T\}, \quad l = 1, \ldots, n,$$

and we know that the corresponding solution $T$ of (1.2) is not unique.

In many situations, there exist conjugate pairs (1.1) and (1.2) such that their solutions are equivalent. However, a system that has the form of (1.2) does not necessarily correspond to a minimization problem as defined in (1.1). With some regularity conditions which will be stated in Chapter 3, we focus on the solution of a system that has the form of (1.2). Our choice is not arbitrary since specific assumptions on the score functions are required in our approximation.

The $M$-estimator is also known as the generalized maximum likelihood estimator. The name comes from the fact that if we choose

$$\tau_l(Y_l, \eta) = -\log\{f_l(Y_l, \eta)\}, \quad l = 1, \ldots, n,$$

or equivalently,

$$\Psi_l(Y_l, \eta) = -\frac{1}{f_l(Y_l, \eta)} \cdot \frac{\partial f_l(Y_l, \eta)}{\partial \eta}, \quad l = 1, \ldots, n,$$

the corresponding $M$-estimator can be the ordinary maximum likelihood estimator. In other words, the $M$-estimator is a general class of estimators which includes some maximum likelihood estimators as special cases. We want to know if the general estimator possesses the nice properties of the special one.

Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed, $\hat{\eta}_n$ is any sequence of functions such that

$$\frac{1}{n} \sum_{l=1}^{n} \Psi_l(Y_l, \hat{\eta}_n) \to 0 \tag{1.3}$$

almost surely (or in probability), where $\Psi_1 = \cdots = \Psi_n$. Huber gave sufficient conditions for the following two results to hold. The conditions generally require that the function $\Psi_1(Y_1, \eta)$ satisfies certain continuity properties, and the expected value $E[\Psi_1(Y_1, \eta)]$ exists for all $\eta \in \Omega$ and has a unique zero at $\eta = \eta_0$.

**Theorem 1.1** (Huber, 1981, page 132) *Every sequence $\hat{\eta}_n$ satisfying* (1.3) *converges to $\eta_0$ almost surely. An analogous statement is true for convergence in probability.*

□

**Theorem 1.2** (Huber, 1981, page 133) *$\sqrt{n}(\hat{\eta}_n - \eta_0)$ is asymptotically normal with mean 0 and covariance matrix $A^{-1}C(A^T)^{-1}$, where $A$ is the nonsingular derivative matrix of $E[\Psi_1(Y_1, \eta)]$ at $\eta = \eta_0$ and $C$ is the covariance matrix of $\Psi_1(Y_1, \eta_0)$.*

□

The above results have been extended to various situations and in particular to regression problems where the random observations are not identically distributed. We are not going to restate them here. However, specific results will be given when we use them in the examples.

In addition to the asymptotic results above, the $M$-estimator has other important features. In practice, we want our estimators to be robust in some sense. The flexibility in the choice of the score function for an $M$-estimator allows us to define an estimator which satisfies some prescribed properties. We illustrate the idea through a simple location problem from Huber (1964).

Let $Y_1, \ldots, Y_n$ be independent and share a common density function $f_{\mu_0}(y)$. Suppose that we want to define an estimator for the location parameter $\mu_0$, which resists outliers but at the same time retains a high efficiency. We can choose Huber's score function with a specific $c$,

$$\Psi_c(r) = max\{-c, min\{c, r\}\},$$

and solve the equation

$$\sum_{l=1}^{n} \Psi_c(Y_l - \hat{\mu}) = 0.$$

Huber showed that

$$\hat{\mu} \rightsquigarrow N\left(\mu_0, \frac{E_f[\Psi_c^2(r)]}{n\{E_f[\Psi_c'(r)]\}^2}\right),$$

where '$\rightsquigarrow$' means 'is asymptotically distributed as' and $r = Y_1 - \mu_0$. Note that the estimator includes the sample mean ($c = \infty$) and the sample median ($c = 0$) as the limiting cases. Huber also showed that the estimator has many desired robust

properties. The trade off between asymptotic efficiency and the resistance to outliers is regulated by the choice of $c$. For instance. when the population is normally distributed, the asymptotic efficiency ranges from 1 for $c = \infty$ to .64 for $c = 0$. A typical choice for $c$ is 1.345 which corresponds to a .95 asymptotic efficiency at the normal.

Another advantage of the $M$-estimator over other classes of estimators is that its definition can easily be extended from one-parameter to multi-parameter and from univarate to multivariate problems. On the contrary, an estimator based on the rank or the order statistics generally suffers from the difficulty in ordering for multivariate cases.

We make a final remark even though it is not particularly tied to the $M$-estimators. For applications in robust statistics, the score functions are generally bounded in order to limit the influence of individual observations. It turns out that the boundedness has an additional advantage for our approximation. In brief, we will require the existence of some moment generating functions for our applications. Having bounded score functions guarantees the existence of the moment generating functions.

## 1.3 Exponential tilt

A key to our approximation is the idea of recentering. In brief, we need to transform a density function such that the new density function satisfies some prescribed conditions. A typical condition in statistical applications is to enforce a certain expectation under the new density function. The details of this will be presented in Chapter 3. In this section, we introduce a technique called the exponential tilt that we use for the recentering.

The exponential tilt has widely been used in statistics and especially in the area of small sample asymptotics (see Field and Ronchetti, 1990). In particular, Field (1982) derived an approximation for the joint density function of a multivariate $M$-estimator by applying an exponential tilt at each point where the density is to be approximated. More examples of its applications will be given in the next chapter.

To develop the ideas, let $f(y)$ be the density function of a random variable $Y$. The moment generating function of $Y$ under $f$ is defined by

$$M_f(\alpha) = E_f[exp\{\alpha Y\}] = \int_y exp\{\alpha y\}\ f(y)\ dy,$$

where $\alpha$ is real. We know that $M_f(\alpha)$ does not exist for all $\alpha$ and $Y$, which can occur even when $Y$ has a commonly used distribution such as a $t$ distribution. This indeed causes some problems in applying the exponential tilt. Nevertheless, the existence of the moment generating function is guaranteed when $Y$ is bounded. In our applications, the role of $Y$ is taken by a score function, which is generally bounded for robust $M$-estimators and the existence problem disappears.

When $M_f(\alpha)$ exists for a given $\alpha_0$, a conjugate or exponentially tilted density function of $Y$ for the given $\alpha_0$ is defined by

$$h(y) = c(\alpha_0)\ exp\{\alpha_0 y\}\ f(y),$$

where

$$c^{-1}(\alpha_0) = \int_y exp\{\alpha_0 y\}\ f(y)\ dy = M_f(\alpha_0).$$

We will call $h(y)$ the $\alpha_0$-conjugate density function of $f(y)$. It follows from the definitions of the moment generating function and the $\alpha_0$-conjugate density function that

$$M_h(\alpha) = \frac{M_f(\alpha + \alpha_0)}{M_f(\alpha_0)}.$$

In addition, we have the following simple but heuristic result.

**Lemma 1.1** *$h(y)$ is the $\alpha_0$-conjugate density function of $f(y)$ if and only if $f(y)$ is the $(-\alpha_0)$-conjugate density function of $h(y)$.*

**Proof** Let $h(y)$ be the $\alpha_0$-conjugate density function of $f(y)$, and $g(y)$ be the $(-\alpha_0)$-conjugate density function of $h(y)$. By definition,

$$g(y) = d(-\alpha_0)\ exp\{-\alpha_0 y\}\ h(y) = d(-\alpha_0)\ c(\alpha_0)\ f(y),$$

where

$$d^{-1}(-\alpha_0) = \int_y c(\alpha_0)\ f(y)\ dy = c(\alpha_0).$$

Therefore we have $g(y) = f(y)$. This proves the 'only if' part, the 'if' part is similar and is omitted.

□

We notice that a moment generating function does not always exist. However, if it exists, the moment generating function is unique and completely determines the distribution of the random variable (see Hogg and Craig, 1978, page 50). In other words, we can study the characteristics of a random variable via its moment generating function. The above lemma implies that if $h(y)$ is the $\alpha_0$-conjugate density function of $f(y)$, then

$$M_f(\alpha) = \frac{M_h(\alpha - \alpha_0)}{M_h(-\alpha_0)}.$$

This suggests an indirect alternative approach to understanding the properties of a random variable under its original density function. That is, we can study its behaviour under a conjugate density function and transform the result back through the connection of the two moment generating functions. The idea had been applied in Field (1982). We will utilize it into a more general situation.

We have discussed an existence problem of the conjugate density functions. In addition to the existence of $h(y)$, we need the next two results from Daniels in order to satisfy some required conditions in our approximation.

Let $F(y)$ be the distribution function of $Y$, and define $K(\alpha) = log(M(\alpha))$. Note that $K(\alpha)$ is called the cumulant generating function of $Y$.

**Theorem 1.3** (Daniels, 1954) *$F(y) = 0$ for $y < a$, and $F(y) = 1$ for $y > b$ if and only if $K(\alpha)$ exists for all real $\alpha$ and $K'(\alpha) = t$ has no real root whenever $t < a$ or $t > b$.*

□

**Theorem 1.4** (Daniels, 1954) *Let $F(y) = 0$ for $y < a$, $0 < F(y) < 1$ for $a < y < b$, $F(y) = 1$ for $y > b$, where $-\infty < a < b < \infty$. Then for every $t_0$ in $a < t_0 < b$ there*

*is a unique simple real root* $\alpha_0$ *of* $K'(\alpha_0) = t_0$. *As* $\alpha$ *increases from* $-\infty$ *to* $\infty$, $K'(\alpha)$ *increases continuously from* $t = a$ *to* $t = b$.

□

For our approximation, recentering and manipulating a connection between the moment generating functions $M_h(y)$ and $M_f(y)$ are the two major tools. The exponential tilt allows us to recenter the original density function $f(y)$ to a new density function $h(y)$ that satisfies some required conditions. Moreover, it supports a nice relationship between $M_h(y)$ and $M_f(y)$ and eventually the density function of an estimator under $h(y)$ and that under $f(y)$. We will see in Chapter 3 how the exponential tilt enables us to focus our problem on the approximation of the density at the expected value of an estimator. This is important since the Edgeworth approximation for densities (see Section 1.4) generally provides very accurate numerical results around the expected value.

The primary advantage of using the exponential tilt in our approximation comes perhaps not from its theoretical properties but rather from the functional form of a conjugate density function. It is the exponential form in its definition that allows us to derive a relationship between the two density functions of an estimator under $h(y)$ and $f(y)$. In addition, we will see in the development that the form also allows us to simplify the relationship by eliminating a messy conditional expectation.

Nevertheless, there are theoretical justifications for the exponential tilt. For instance, Tingley and Field (1990) discussed the issue based on the results in Kullback (1959), and concluded that the exponential tilt forces $h(y)$ to satisfy some prescribed conditions while altering $f(y)$ as little as possible in the Kullback-Leibler distance. However, this is not essential to our approximation and we will not go into the details.

## 1.4 Approximation of the mean

In our approximation, we are required to approximate the density at the expected value of the arithmetic mean of $n$ independent functions, where the functions are not necessarily identically distributed. To achieve that, we apply a local normal approximation. In this section, we give the approximation and state some results which ensure the accuracy of the approximation.

Let $S_n$ be the sum of $n$ independent random variables $Y_1, \ldots, Y_n$. Discussions on the conditions for $S_n$ to be asymptotically normally distributed can be found in numerous works. In particular, we state a result due to Liapunov (see Prakasa Rao, 1987, page 22) for which the conditions are similar to those of Theorem 1.6 that we need.

**Theorem 1.5** (Liapunov) *If* $\{Y_n, n \geq 1\}$ *are independent random variables with* $E[Y_n] = 0$ *and if*

$$\frac{1}{\Upsilon_n^{\delta}} \sum_{l=1}^{n} E|Y_l|^{\delta} \to 0 \quad as \quad n \to \infty \tag{1.4}$$

*for some* $\delta > 2$, *where* $\Upsilon_n^2 = \sigma_1^2 + \cdots + \sigma_n^2$ *and* $\sigma_n^2 = E[Y_n^2] < \infty$, *then*

$$\frac{S_n}{\Upsilon_n} \xrightarrow{L} N(0,1).$$

□

As we mentioned earlier, the roles of $Y_l$'s are taken by the score functions which are generally bounded in our approximation. Therefore the following special case is particularly useful for us.

**Corollary 1.1** *If* $\{Y_n, n \geq 1\}$ *are independent random variables with* $E[Y_n] = 0$ *and if there exist positive numbers* $\varepsilon$ *and* $\varepsilon_\delta$ *such that*

$$0 < \varepsilon \leq \sigma_l^2 < \infty \quad \text{and} \quad E|Y_l|^\delta < \varepsilon_\delta, \quad l = 1, \ldots, n,$$

*for some* $\delta > 2$, *then*

$$\frac{S_n}{\Upsilon_n} \xrightarrow{L} N(0,1).$$

**Proof** Since the boundedness implies that

$$\Upsilon_n^2 \geq n\varepsilon \quad \textit{and} \quad \sum_{l=1}^{n} E|Y_l|^\delta < n\varepsilon_\delta,$$

we have

$$0 \leq \frac{1}{\Upsilon_n^\delta} \sum_{l=1}^{n} E|Y_l|^\delta < \frac{n\varepsilon_\delta}{(n\varepsilon)^{\frac{\delta}{2}}} \to 0$$

as $n \to \infty$. Therefore the condition (1.4) is satisfied and the result follows.

□

Suppose that $S_n$ is asymptotically normal. Then, when the exact distribution of $S_n$ is not available, one may want to use the asymptotic result and hope that it will give a reasonable approximation to the exact distribution. To measure the quality of such an approximation, we need to know the error induced by the approximation. For this purpose, we have the following result.

**Theorem 1.6** (Esseen, 1945, page 43) *Let* $Y_1, Y_2, \ldots, Y_n$ *be a sequence of independent random variables such that each variable* $Y_l$ *has mean value zero and the finite absolute moment* $\beta_{\delta l}$ *of given order* $\delta$, $2 < \delta \leq 3$. *Then*

$$\left| P\left\{ \frac{S_n}{\Upsilon_n} < y \right\} - \Phi(y) \right| \leq c_\delta \left( \frac{\xi_{\delta n}}{n^{\frac{\delta-2}{2}}} + \frac{\xi_{\delta n}^{\frac{1}{\delta-2}}}{n^{\frac{1}{2}}} \right),$$

*where*

$$\xi_{\delta n} = \frac{B_{\delta n}}{B_{2n}^{\frac{\delta}{2}}}, \quad B_{\delta n} = \frac{1}{n}\sum_{l=1}^{n} \beta_{\delta l},$$

*$c_\delta$ is a finite positive constant depending only on $\delta$, and $\Phi(\cdot)$ is the standard normal cumulative distribution function.*

□

We observe that the error bound of the normal approximation is of order $O\left(n^{-\frac{1}{2}}\right)$ and cannot be improved in general. Different approaches have been proposed to increase the accuracy. We shall see some of them in the next chapter. Now, we discuss the idea that we use in our approximation.

We realize that the normal approximation generally works well around the expected value of an arithmetic mean and can be very inaccurate in the tails. Our philosophy is simply to use only the best! With an application of the exponential tilt, we will see that all we need is a good approximation at the expected value, the place where a normal approximation generally gives satisfactory results. This idea can be verified by a formal Edgeworth expansion. We state a result of Feller.

**Theorem 1.7** (Feller, 1971, page 535, see also Field and Ronchetti, 1990, page 11) *Let $Y_1, \ldots, Y_n$ be $n$ independent and identically distributed random variables with distribution function $F$ and characteristic function $\psi$. Let*

$$E[Y_l] = \mu_1 = 0, \quad var(Y_l) = \sigma^2 < \infty, \quad l = 1, \ldots, n,$$

*and*

$$F_n(t) = P\left\{\frac{1}{\sqrt{n}}\sum_{l=1}^{n}\frac{Y_l}{\sigma} < t\right\}$$

*with density $f_n(t)$. Suppose that the moments $\mu_3, \ldots, \mu_k$ exist and that $|\psi|^\nu$ is integrable for some $\nu > 1$. Then, $f_n$ exists for $n > \nu$ and as $n \to \infty$,*

$$f_n(t) - \varphi(t) - \varphi(t) \sum_{r=3}^{k} \frac{P_r(t)}{n^{\frac{r}{2}-1}} = o\left(\frac{1}{n^{\frac{k}{2}-1}}\right), \tag{1.5}$$

*uniformly in $t$. Here $P_r$ is a real polynomial of degree $3(r-2)$ depending only on $\mu_1, \ldots, \mu_r$ but not on $n$ and $k$ (or otherwise on $F$), and $\varphi$ is the standard normal density function.*

□

The expansion (1.5) is called the Edgeworth expansion of $f_n$. When $k = 4$, Feller showed that

$$f_n(t) - \varphi(t) - \frac{1}{\sqrt{n}} P_3(t)\varphi(t) - \frac{1}{n} P_4(t)\varphi(t) = o\left(\frac{1}{n}\right),$$

where

$$P_3 = \frac{\mu_3}{6\sigma^3} H_3, \quad P_4 = \frac{\mu_3^2}{72\sigma^6} H_6 + \frac{\mu_4 - 3\sigma^4}{24\sigma^4} H_4,$$

and

$$H_3(t) = t^3 - 3t, \quad H_4(t) = t^4 - 6t^2 + 3, \quad \text{and} \quad H_6(t) = t^6 - 15t^4 + 45t^2 - 15$$

are the Hermite polynomials of order 3, 4, and 6 respectively. Again, the one-term normal approximation gives an error of order $O\left(n^{-\frac{1}{2}}\right)$. However, when $t = 0$, $H_3(t) = 0$ and we obtain

$$f_n(0) - \frac{1}{\sqrt{2\pi}} = O\left(\frac{1}{n}\right). \tag{1.6}$$

This result is a key to the high accuracy of our approximation.

Since the first time Edgeworth (1905) derived the expansion, similar expansions have been developed under various conditions such as $Y_l$'s not being identically distributed or being multivariate random variables. A basic result on multivariate Edgeworth expansions is given in Bhattacharya and Ghosh (1978). More results and references can be found in Hall (1992).

For our development in Chapter 3, we use a local approximation from the result in (1.6). Precisely, let $Y = (Y_1, \ldots, Y_n)$ be an independent random sample such that $E[Y_l] = 0$ and $0 < var(Y_l) < \infty$. Define $\overline{Y}$ to be the sample mean. To approximate the density of $\overline{Y}$ at zero, we use the normal approximation

$$f_{\overline{Y}}(0) = \frac{1}{\sqrt{2\pi\sigma^2_{\overline{Y}}}} + O\left(\frac{1}{n}\right),$$

where $\sigma^2_{\overline{Y}}$ is the variance of $\overline{Y}$. Note that a general result for a multivariate mean is also available (see McCullagh, 1987, page 150). We will state the general result when it is needed in Chapter 5.

## 1.5 Summary

The purpose of the present work is to derive an accurate approximation for the marginal densities of a multivariate $M$-estimator. The result is particularly useful in robust statistics where an estimator usually does not have a closed form, and for small sample applications where the asymptotic results may not be reliable. We summarize the general problem as follows.

Let $Y = (Y_1, \ldots, Y_n)$ be an independent random sample, where each of the $Y_l$'s has a density function $f_l(y_l)$ that is parameterized by $\eta = (\eta_1, \ldots, \eta_p)$. Let $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_p)$ be an $M$-estimator of $\eta \in \Omega \subset \Re^p$, that is, the solution of a system of $p$ equations

$$\frac{1}{n}\sum_{l=1}^{n} \Psi_l(Y_l, \hat{\eta}) = 0,$$

where $\Psi_l = \{\Psi_{1l}, \ldots, \Psi_{pl}\}$. Our primary objective is to derive an approximation for the marginal density of a component in $\hat{\eta}$ under $\eta = \eta_0$. The result is then extended to a real-valued function $\rho(\hat{\eta})$, $\rho : \Re^p \to \Re$, and finally to a real-valued vector $\rho(\hat{\eta}) = \{\rho_1(\hat{\eta}), \ldots, \rho_k(\hat{\eta})\}$, $\rho : \Re^p \to \Re^k$, $k \leq p$.

In this chapter, we have given an overview to the general problem. In particular, we have discussed the importance of the $M$-estimator and the need of an accurate approximation for its finite sample behaviour. We have derived the basic philosophy for our approximation and introduced the exponential tilt and a local density approximation. The discussion is accompanied by the theoretical results which are needed in our development.

Since the $M$-estimator was proposed by Huber in 1964, different techniques have been developed directly or indirectly to approximate its distribution. In the next chapter, we give a brief account on some of the recent work which is closely related to

our approximation. Our main results and their derivations are presented in Chapter 3. In Chapter 4, we apply our results to several linear regression problems, and some numerical results are generated for comparison. Chapter 5 extends the results to a real-valued vector and gives numerical examples to verify the accuracy. Finally in Chapter 6 we summarize our results and give some concluding remarks. The technical details for the computation of our approximation are given in the appendices.

# Chapter 2

# Related techniques

## 2.1 Overview

In Chapter 1, we defined the general problem of interest. We now begin to find a solution for it. In this chapter, we review several existing techniques which are related to our problem and discuss some of their features.

To approximate the marginal distribution of an $M$-estimator, we have basically two different approaches which can be referred to as the large sample and the small sample methods. The former one solves the problem via some asymptotic results while the latter one works directly on the finite sample behaviour. Generally speaking, the first approach is simpler but the second one is more accurate.

When the asymptotic distribution of an estimator is known, we can use it to approximate the finite sample distribution. If the accuracy is not good enough, an option is to modify the asymptotic result to improve the approximation. On the other hand, when such a result is not available or its performance for a small sample is unclear, an alternative is to derive directly a finite sample approximation. For the finite sample approach, there exist at least two immediate options. The first is to

work on the distribution function and the second is to approximate the estimator itself.

In the past three decades, various results using the different approaches have been derived. For instance, Huber (1964, 1973) obtained many results describing the asymptotic behaviour of $M$-estimators under several different conditions and gave sufficient conditions for the results to hold. Field (1982) developed an approximation for the joint density function of a multivariate $M$-estimator, and Tingley and Field (1990) derived a linear approximation for a real-valued function of a multivariate $M$-estimator.

In the next section, we introduce two models and two estimators to illustrate our discussion. They will be employed in different sections of this chapter. Section 2.3 presents some asymptotic results for the estimators. Sections 2.4 and 2.5 discuss the work of Field (1982) and Tingley and Field (1990) respectively. In particular, the work by Field requires a multi-dimensional integration to obtain a marginal distribution. At this point, DiCiccio and Martin (1991) provide a useful approximation which allows us to avoid a numerical integration. Their approximation is given in Section 2.6. Finally in Section 2.7 we summarize and compare the different techniques.

Before beginning the next section, we make several remarks on our discussion in Sections 2.3 to 2.6. For the different techniques, we will state only the main results. The notation is unified for convenient comparison. The developments and underlying assumptions will not be reproduced unless they are related to our discussion. Nevertheless, references will always be given for the details.

## 2.2 Two models and their estimators

In this section we define two statistical models, and two $M$-estimators $\hat{\eta}$ for their parameters $\eta$. They are not directly related to the techniques that we are going to discuss, but it is useful to have them in the discussion. In particular, the asymptotic results that we will present in the next section are based on these models and estimators. Moreover, they will help us to adopt the general notation in our approximation, and will be the basis of the examples in Chapters 4 and 5. However, one should realize that the applications of the different techniques in our discussion are not restricted to these models and estimators.

To begin, let $Y = (Y_1, \ldots, Y_n)$ be an independent random sample of size $n$.

The first model is a location-scale model, that is,

$$Y_l = \theta + \sigma\varepsilon_l, \quad l = 1, \ldots, n,$$

where $\theta$ is a location parameter, $\sigma > 0$ is a scale parameter, and $\varepsilon_l$'s are independent and identically distributed random errors having the common density function $f_e$. Therefore we have $\eta = (\theta, \sigma)$, and independent and identically distributed $Y_l$'s with the common density function

$$f(y_l) = \frac{1}{\sigma} f_e\left(\frac{y_l - \theta}{\sigma}\right).$$

For an estimator of $\eta$, we choose Huber's proposal 2 (Huber, 1964) in which the score functions are

$$\Psi_{1l}(Y_l, \eta) = \Psi_c\left(\frac{Y_l - \theta}{\sigma}\right) \quad \text{and} \quad \Psi_{2l}(Y_l, \eta) = \Psi_c^2\left(\frac{Y_l - \theta}{\sigma}\right) - \beta,$$

$l = 1, \ldots, n$, where $\Psi_c$ is the Huber's score function defined in Section 1.2, and $\beta$ is a constant to be specified.

The second model is a multiple regression model, that is,

$$Y = X\theta + \sigma\varepsilon,$$

where $X$ is an $n$ by $p-1$ fixed design, $\theta = (\theta_1, \ldots, \theta_{p-1})$, $\sigma > 0$ is a scale parameter, and $\varepsilon$ is an $n$-vector of independent and identically distributed random errors. In this model, $\eta = (\theta, \sigma)$ is $p$-dimensional, and the random observations $Y_l$'s are independent but not necessarily identically distributed. We denote the density functions of $Y_l$ and $\varepsilon_l$ by $f_l$ and $f_\varepsilon$ respectively. Hence,

$$f_l(y_l) = \frac{1}{\sigma} f_\varepsilon \left( \frac{y_l - X_l^T \theta}{\sigma} \right).$$

The joint density function of $Y_l$'s is denoted by $f$, that is,

$$f(y) = \prod_{l=1}^{n} f_l(y_l).$$

To estimate the parameter $\eta$, we use Huber-type score functions

$$\Psi_{jl}(Y_l, \eta) = \Psi_c \left( \frac{Y_l - X_l^T \theta}{\sigma} \right) X_{lj}, \quad j = 1, \ldots, p-1,$$

and

$$\Psi_{pl}(Y_l, \eta) = \Psi_c^2 \left( \frac{Y_l - X_l^T \theta}{\sigma} \right) - \beta,$$

$l = 1, \ldots, n$, where $\Psi_c$ is the Huber's score function and $\beta$ is a constant. Note that the least squares estimator is the limiting case corresponding to $c = \infty$.

The two models are chosen mainly because of their popularity in practice. We realize that the first model is in fact included in the second one. However, the location-scale model has its own importance for identically distributed variables and

a simpler notation. For both models, we assume that the true value of the parameter is $\eta_0 = (\theta_0, \sigma_0)$.

The Huber-type estimators for the two models are denoted by $\hat{\eta} = (\hat{\theta}, \hat{\sigma})$ with the difference being that $\hat{\theta}$ is a scalar for the location-scale model and a $(p-1)$-vector for the regression model. We use them in our discussion for several reasons. The Huber-type estimator has been central to many recent developments of robust procedures for the two models, and has some desired properties. For instance, it is asymptotically normal under some general conditions, so we are able to compute a large sample approximation for comparisons. Furthermore, the Huber-type estimator satisfies the natural invariance requirements for estimators in a regression context.

The idea of location-scale equivariance and invariance for an estimator appears in numerous works (see Lawless, 1982, page 538; Staudte and Sheather, 1990, page 101). For the multiple regression model the following definition gives the natural invariance requirements. Note that we put additional subscripts to emphasize the dependence on the parameters.

**Definition 2.1** *Consider a multiple regression model* $Y_{l,\theta,\sigma} = X_l^T\theta + \sigma\varepsilon_l$, $l = 1, \ldots, n$, *where* $\theta$ *is a* $(p-1)$*-vector and* $\sigma > 0$. *Denote the joint distribution function of* $Y_{l,\theta,\sigma}$*'s by* $F_{\theta,\sigma}$. *Suppose that* $(\hat{\theta}_{\theta,\sigma}, \hat{\sigma}_{\theta,\sigma})$ *is an estimator of the p-dimensional parameter* $(\theta, \sigma)$ *under* $F_{\theta,\sigma}$. *The estimator* $\rho(\hat{\theta}_{\theta,\sigma}, \hat{\sigma}_{\theta,\sigma})$ *of a parametric function* $\rho(\theta_{\theta,\sigma}, \sigma_{\theta,\sigma})$ *is called location equivariant if* $\rho(\hat{\theta}_{\theta+b,\sigma}, \hat{\sigma}_{\theta+b,\sigma}) = \rho(\hat{\theta}_{\theta,\sigma}, \hat{\sigma}_{\theta,\sigma}) + b$, *or location invariant if* $\rho(\hat{\theta}_{\theta+b,\sigma}, \hat{\sigma}_{\theta+b,\sigma}) = \rho(\hat{\theta}_{\theta,\sigma}, \hat{\sigma}_{\theta,\sigma})$, *for all* $(p-1)$*-vectors b. It is called scale equivariant if* $\rho(\hat{\theta}_{a\theta,a\sigma}, \hat{\sigma}_{a\theta,a\sigma}) = a\rho(\hat{\theta}_{\theta,\sigma}, \hat{\sigma}_{\theta,\sigma})$, *or scale invariant if* $\rho(\hat{\theta}_{a\theta,a\sigma}, \hat{\sigma}_{a\theta,a\sigma}) = \rho(\hat{\theta}_{\theta,\sigma}, \hat{\sigma}_{\theta,\sigma})$, *for all* $a > 0$.

□

**Theorem 2.1** *The Huber-type estimators* $(\hat{\theta}_{\theta,\sigma}, \hat{\sigma}_{\theta,\sigma})$ *defined for our models are scale equivariant. In addition,* $\hat{\theta}_{\theta,\sigma}$ *is location equivariant, and* $\hat{\sigma}_{\theta,\sigma}$ *is location invariant.*

**Proof** We need only to consider the multiple regression estimator. Recall that the score functions of the Huber-type estimator are in the form of

$$\Psi_{jl}(Y_{l,\theta,\sigma}, \theta, \sigma) = \Psi_{jl}\left(\frac{Y_{l,\theta,\sigma} - X_l^T\theta}{\sigma}\right), \quad j = 1, \ldots, p, \quad l = 1, \ldots, n,$$

for any design matrix $X$. Therefore by the definition of $(\hat{\theta}_{\theta,\sigma}, \hat{\sigma}_{\theta,\sigma})$, we have

$$\sum_{l=1}^{n} \Psi_{jl}\left(\frac{Y_{l,\theta,\sigma} - X_l^T\hat{\theta}_{\theta,\sigma}}{\hat{\sigma}_{\theta,\sigma}}\right) = 0. \tag{2.1}$$

Let $(\hat{\theta}_{a\theta+b,a\sigma}, \hat{\sigma}_{a\theta+b,a\sigma})$ be the estimator under $F_{a\theta+b,a\sigma}$. Then

$$\sum_{l=1}^{n} \Psi_{jl}\left(\frac{Y_{l,a\theta+b,a\sigma} - X_l^T\hat{\theta}_{a\theta+b,a\sigma}}{\hat{\sigma}_{a\theta+b,a\sigma}}\right) = 0. \tag{2.2}$$

Since

$$Y_{l,a\theta+b,a\sigma} \;=\; X_l^T(a\theta + b) + a\sigma\varepsilon_l \;=\; aY_{l,\vartheta,\sigma} + X_l^T b,$$

the system (2.2) implies that

$$\sum_{l=1}^{n} \Psi_{jl}\left(\frac{Y_{l,\theta,\sigma} - X_l^T\dfrac{\hat{\theta}_{a\theta+b,a\sigma} - b}{a}}{\dfrac{\hat{\sigma}_{a\theta+b,a\sigma}}{a}}\right) = 0. \tag{2.3}$$

Comparing (2.1) and (2.3) gives

$$\hat{\theta}_{\theta,\sigma} = \frac{\hat{\theta}_{a\theta+b,a\sigma} - b}{a} \quad \text{and} \quad \hat{\sigma}_{\theta,\sigma} = \frac{\hat{\sigma}_{a\theta+b,a\sigma}}{a},$$

which imply in particular that

$$\hat{\theta}_{a\theta,a\sigma} = a\hat{\theta}_{\theta,\sigma}, \quad \hat{\sigma}_{a\theta,a\sigma} = a\hat{\sigma}_{\theta,\sigma}, \quad \hat{\theta}_{\theta+b,\sigma} = \hat{\theta}_{\theta,\sigma} + b, \quad \hat{\sigma}_{\theta+b,\sigma} = \hat{\sigma}_{\theta,\sigma}.$$

□

**Corollary 2.1** *The function*

$$\rho(\hat{\theta}_{\theta,\sigma}, \hat{\sigma}_{\theta,\sigma}) = \rho\left(\frac{\hat{\theta}_{\theta,\sigma} - \theta}{\hat{\sigma}_{\theta,\sigma}}\right)$$

*is location and scale invariant, or simply location-scale invariant.*

**Proof** It follows from Theorem 2.1 that

$$\begin{aligned}
\rho(\hat{\theta}_{a\theta+b,a\sigma}, \hat{\sigma}_{a\theta+b,a\sigma}) &= \rho\left(\frac{\hat{\theta}_{a\theta+b,a\sigma} - (a\theta + b)}{\hat{\sigma}_{a\theta+b,a\sigma}}\right) \\
&= \rho\left(\frac{(a\hat{\theta}_{\theta,\sigma} + b) - (a\theta + b)}{a\hat{\sigma}_{\theta,\sigma}}\right) \\
&= \rho(\hat{\theta}_{\theta,\sigma}, \hat{\sigma}_{\theta,\sigma}).
\end{aligned}$$

□

The last result is practically useful because it allows us to create location-scale invariant statistics for inferential purposes. This idea will be elaborated in Chapter 4. From now on, the extra subscripts on the estimators will be dropped. Unless specified otherwise, the settings of these models and estimators will be clear from the discussion.

## 2.3 Asymptotic distributions

When a new class of estimators is proposed, its general behaviour should be investigated. Very often, the asymptotic distribution is relatively easy to derive through, possibly, a central limit theorem. We can then use the result to obtain some asymptotic properties of the estimator, and when the finite sample distribution is not available, use it as a natural approximation.

We discussed in Chapter 1 several asymptotic results for the $M$-estimator. In particular, Huber showed that under some general conditions (Huber, 1981, pages 131, 132), the estimator is asymptotically normally distributed. We here demonstrate the results through a particular application.

Consider the Huber-type estimator for our location-scale model. Define

$$r_1 = \frac{Y_1 - \theta}{\sigma}.$$

Recall that the score function for the estimator is defined as

$$\Psi_1(Y_1, \eta) = \left\{ \begin{array}{c} \Psi_c(r_1) \\ \Psi_c^2(r_1) - \beta \end{array} \right\}.$$

Therefore the matrices $A$ and $C$ in Theorem 1.2 are given by

$$\begin{aligned} A &= E_f \left[ \left. \frac{\partial \Psi_1(Y_1, \eta)}{\partial \eta^T} \right|_{\eta=\eta_0} \right] \\ &= -\frac{1}{\sigma_0} E_f \left[ \begin{array}{cc} I_c(\varepsilon_1) & \varepsilon_1 I_c(\varepsilon_1) \\ 2\varepsilon_1 I_c(\varepsilon_1) & 2\varepsilon_1^2 I_c(\varepsilon_1) \end{array} \right] \end{aligned}$$

and

$$\begin{aligned} C &= E_f[\Psi_1(Y_1, \eta_0)\Psi_1^T(Y_1, \eta_0)] \\ &= E_f \left[ \begin{array}{cc} \Psi_c^2(\varepsilon_1) & \Psi_c(\varepsilon_1)(\Psi_c^2(\varepsilon_1) - \beta) \\ \Psi_c(\varepsilon_1)(\Psi_c^2(\varepsilon_1) - \beta) & (\Psi_c^2(\varepsilon_1) - \beta)^2 \end{array} \right], \end{aligned}$$

where

$$I_c(x) = \begin{cases} 1, & \text{if } |x| < c \\ 0, & \text{otherwise} \end{cases}.$$

If we assume that the random errors are symmetrically distributed about zero, the two matrices simplify to

$$A = -\frac{1}{\sigma_0}\begin{Bmatrix} E_f[I_c(\varepsilon_1)] & 0 \\ 0 & 2E_f[\varepsilon_1^2 I_c(\varepsilon_1)] \end{Bmatrix}$$

and

$$C = \begin{Bmatrix} E_f[\Psi_c^2(\varepsilon_1)] & 0 \\ 0 & E_f[(\Psi_c^2(\varepsilon_1) - \beta)^2] \end{Bmatrix}.$$

So we have

$$A^{-1} = (A^T)^{-1} = -\sigma_0\begin{Bmatrix} \{E_f[I_c(\varepsilon_1)]\}^{-1} & 0 \\ 0 & \{2E_f[\varepsilon_1^2 I_c(\varepsilon_1)]\}^{-1} \end{Bmatrix}$$

and

$$A^{-1}C(A^T)^{-1} = \sigma_0^2\begin{Bmatrix} \dfrac{E_f[\Psi_c^2(\varepsilon_1)]}{\{E_f[I_c(\varepsilon_1)]\}^2} & 0 \\ 0 & \dfrac{E_f[(\Psi_c^2(\varepsilon_1) - \beta)^2]}{4\{E_f[\varepsilon_1^2 I_c(\varepsilon_1)]\}^2} \end{Bmatrix}.$$

It follows from Theorem 1.2 that the location and the scale estimators are asymptotically independent. In addition, the asymptotic distribution of the location estimator $\hat{\theta}$ is given by

$$\hat{\theta} \rightsquigarrow N\left(\theta_0, \frac{\sigma_0^2}{n}\frac{E_f[\Psi_c^2(\varepsilon_1)]}{\{E_f[I_c(\varepsilon_1)]\}^2}\right). \tag{2.4}$$

For the multiple regression model, numerous asymptotic results exist for the $M$-estimators under different conditions (see Huber, 1973, Yohai and Maronna, 1979, and Maronna and Yohai, 1981). In particular, Yohai and Maronna (1979) show that under

some general conditions our Huber-type estimator $\hat{\theta}$ has the asymptotic multivariate normal distribution

$$\hat{\theta} \rightsquigarrow N\left(\theta_0, \sigma_0^2 \frac{E_f[\Psi_c^2(\varepsilon_1)]}{\{E_f[I_c(\varepsilon_1)]\}^2}(X^T X)^{-1}\right). \tag{2.5}$$

As we can expect from the definition of the estimator, when $c$ is set to infinity, the above distribution simplifies to

$$N\left(\theta_0, \sigma_0^2 var(\varepsilon_1)(X^T X)^{-1}\right),$$

the asymptotic distribution of the least squares estimator.

The two results (2.4) and (2.5) will be used in the numerical examples for comparisons. In addition, classical results suggest the possibility of replacing the normal distribution by a $t$ distribution for a better small sample approximation. We will try this in our examples.

## 2.4 Approximation for joint densities

Approximating the finite sample behaviour of a general $M$-estimator is not a trivial task. The boundedness of a score function improves the robustness of an estimator but at the same time makes it impossible to solve the problem analytically. Different approaches have been attempted to approximate the exact distribution. In particular, an important result was found by Field (1982) which we now present.

Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed random observations from an underlying density function $f_\eta$, where $\eta$ is a $p$-dimensional parameter, $Y_l$'s may be univariate or multivariate. Field derived an approximation for the joint density function of a multivariate $M$-estimator $\hat{\eta}$ of $\eta$ under some regularity conditions. Note that the conditions will be generalized to non-identically distributed variables in Chapter 3.

**Theorem 2.2** (Field, 1982) *If $\hat{\eta}$ represents the solution of*

$$\sum_{l=1}^{n} \Psi_j(Y_l, \hat{\eta}) = 0, \quad j = 1, \ldots, p,$$

*then an asymptotic expansion for the density of $\hat{\eta}$, say $g_T(t_0)$, is*

$$g_T(t_0) = \left(\frac{n}{2\pi}\right)^{\frac{p}{2}} c^{-n}(t_0) \left\{ |det\ A||det\ C|^{-\frac{1}{2}} + O\left(\frac{1}{n}\right) \right\},$$

*where*

$$c^{-1}(t_0) = \int_{y_1} exp\left\{ \sum_{j=1}^{p} \alpha_j(t_0) \Psi_j(y_1, t_0) \right\} f(y_1)\, dy_1,$$

*$\alpha(t_0) = \{\alpha_1(t_0), \ldots, \alpha_p(t_0)\}$ is the solution of*

$$\int_{y_1} \Psi_j(y_1, t_0)\, exp\left\{ \sum_{j=1}^{p} \alpha_j(t_0) \Psi_j(y_1, t_0) \right\} f(y_1)\, dy_1 = 0 \quad for \quad j = 1, \ldots, p,$$

$$A = \left\{ E\left[ \left. \frac{\partial \Psi_{j_1}(Y_1, \eta)}{\partial \eta_{j_2}} \right|_{\eta = t_0} \right] \right\}_{1 \le j_1, j_2 \le p},$$

$$C = \{E[\Psi_{j_1}(Y_1, t_0)\Psi_{j_2}(Y_1, t_0)]\}_{1 \le j_1, j_2 \le p},$$

*and all expectations are with respect to the conjugate density*

$$h(y_1) = c(t_0)\, exp\left\{ \sum_{j=1}^{p} \alpha_j(t_0)\Psi_j(y_1, t_0) \right\} f(y_1).$$

*The error term holds uniformly for all* $t_0$ *in a compact set.*

□

The joint density approximation was applied to several cases in Field (1982) with excellent results. In particular, it was shown that when the approximation is applied to a multivariate mean, that is,

$$\Psi_j(Y_l, \eta) = Y_{lj} - \eta_j,$$

for $Y_l = (Y_{l1}, \ldots, Y_{lp})$, the approximation is exact if the underlying density $f_\eta$ is multivariate normal, and is exact up to a constant if $Y_l$'s have a common Wishart density.

The original arguments in the derivation of this joint density approximation require that the random observations to be independent and identically distributed. Field and Ronchetti extended the result to non-identically distributed observations. Specifically, they applied the approximation to a simple regression problem with unknown scale parameter (Field and Ronchetti, 1990, page 72). As a special case, they showed that for the least squares estimator under a univariate normal density, the approximation agrees with the exact density up to the constant of integration.

The ideas in this technique motivate our present work in several aspects. However, for the reasons that we will discuss in Section 2.7, we are not going to implement this approximation for numerical comparisons. Instead, we summarize the approach into the following three-step procedure for general comparisons to our work.

Consider the joint density of $\hat{\eta}$ at $t_0$ under $f_\eta$.

Step 1: A conjugate density function $h_{t_0}$ of $f_\eta$ is computed such that $\hat{\eta}$ is centered at $t_0$ in expectation under $h_{t_0}$.

Step 2: A multivariate Edgeworth expansion at zero is used to approximate the joint density of a Taylor series expansion of $\hat{\eta} - t_0$ under $h_{t_0}$.

Step 3: The joint density approximation of $\hat{\eta}$ at $t_0$ under $h_{t_0}$ is transformed to an approximation under $f_\eta$.

## 2.5 Approximation for estimators

In the last two sections, both a large sample and a small sample approach were given as techniques to solve our problem. In brief, the asymptotic distribution gives a simple but unreliable solution for the Huber-type estimator, while the small sample joint density approximation provides a better solution to the problem, but one which is much more complicated. In fact, the complications generate some computational problems which we will discuss later.

While the accuracy of the joint density approximation encourages us to work on the finite sample behaviour, the computational difficulties suggest that we try to approximate directly the marginal distribution. In fact, there is a solution which has combined these two ideas. In this section, we present the work by Tingley and Field (1990) in which a linear approximation of a general $M$-estimator is given.

Suppose that we have an independent and identically distributed sample $Y_1, \ldots, Y_n$ of $m$-dimensional observations drawn from a population with distribution $F_\eta$ involving a $p$-dimensional parameter $\eta$. The parameter $\eta$ has true value $\eta_0$. Let $\hat{\eta}$ be an $M$-estimator of $\eta_0$, which is the solution of

$$\frac{1}{n}\sum_{l=1}^{n}\Psi(Y_l,\hat{\eta}) = 0,$$

where the score function $\Psi$ is $p$-dimensional. Note that the last system differs from the previous definitions by a factor of $n^{-1}$ on the left hand side, which makes no difference at all to the solution.

Under conditions similar to those in Field (1982), Tingley and Field (1990) showed that

$$(\hat{\eta}-\eta_0)_k = \sum_{j=1}^{p} B_{kj}\overline{\Psi}_j + o_p\left(\frac{1}{\sqrt{n}}\right), \quad k = 1,\ldots,p,$$

where

$$\overline{\Psi}_j = \frac{1}{n}\sum_{l=1}^{n}\Psi_j(Y_l,\eta_0), \quad B = \{B_{kj}\} = -A^{-1},$$

and

$$A = E\left[\left.\frac{\partial\Psi(Y,\eta)}{\partial\eta^T}\right|_{\eta=\eta_0}\right].$$

In addition, they showed in the same paper that for a general real-valued function $\rho(\hat{\eta})$, a linear approximation is

$$\rho(\hat{\eta}) - \rho(\eta_0) = \overline{G} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

where

$$\overline{G} = \frac{1}{n}\sum_{l=1}^{n} G_l,$$

and

$$G_l = G(Y_l,\eta_0) = \Psi^T(Y_l,\eta_0)\ B^T\ \left.\frac{\partial\rho(\eta)}{\partial\eta}\right|_{\eta=\eta_0}.$$

Therefore, we may use the distribution of $\overline{G}$ to approximate the exact distribution of $\rho(\hat{\eta})$. For example, consider $\rho(\hat{\eta}) = \hat{\theta}$ in our location-scale problem. Suppose that the random errors are symmetrically distributed about zero. Then from Section 2.3 we have

$$B = -A^{-1} = \sigma_0\begin{Bmatrix} \{E_f[I_c(\varepsilon_1)]\}^{-1} & 0 \\ 0 & \{2E_f[\varepsilon_1^2 I_c(\varepsilon_1)]\}^{-1} \end{Bmatrix}.$$

Hence,

$$\overline{G} = \frac{1}{n}\sum_{l=1}^{n}\frac{\sigma_0}{E_f[I_c(\varepsilon_1)]}\Psi_c\left(\frac{Y_l-\theta_0}{\sigma_0}\right).$$

In particular, when $c = \infty$, $\hat{\theta}$ is the arithmetic mean $\overline{Y}$ and $\overline{G}$ simplifies to $\overline{Y} - \theta_0$. So the approximation is exact. The derivation of the linear approximation for the regression estimator is similar. Details are given in the examples.

In Tingley and Field (1990), the linear approximation is used as a basis for constructing robust confidence intervals. The intervals for $\rho(\eta)$ are obtained by inverting a test for the hypothesis $H : \rho(\eta) = \rho_0$. Since $\eta_0$ is generally unknown, they used the observed value $\eta_{obs}$ from the sample and compute an initial approximation $\overline{G}_{\eta_{obs}}$ of $\rho(\hat{\eta})$ under $F_{\eta_{obs}}$, and then apply an exponential tilt to force $\overline{G}_{\eta_{obs}}$ to satisfy the hypothesis $H$. An application of this idea to our approximation will be discussed in Chapter 6.

The linear approximation is originally derived for an identically distributed random sample. Tingley (1992) extended the result to a general linear regression model and showed that the error of the approximation is $o_p(n^{-\frac{1}{2}})$. The result is useful to our present work.

## 2.6 Approximation for tail probabilities

When the joint density function of a multivariate estimator exists in a closed form and we are interested in the marginal distribution of one of its components, a possibility is to integrate out the unwanted variables. We agree that this can be very accurate but it could be computationally difficult even for a low dimensional problem. An alternative is to approximate the multiple integral and avoid high dimensional integration.

An example of the latter approach is given in DiCiccio, Field, and Fraser (1990). In that paper, an approximation for the marginal tail probability of a component in a random vector was derived. Later, DiCiccio and Martin (1991) applied the result to an approximation of a marginal density introduced by Tierney, Kass, and Kadane (1989), and developed an approximation to a real-valued function of the components. We now present the main result in DiCiccio and Martin (1991).

Consider a continuous random vector $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_p)$ having probability density function of the form

$$g_{\hat{\eta}}(t) = c\, g_1(t)\, exp\{g_2(t)\}, \quad t = (t_1, \ldots, t_p).$$

Suppose that the funcion $g_2$ attains its maximum value at $t = t_{max}$ and that $\hat{\eta} - t_{max}$ is $O_p(n^{-\frac{1}{2}})$ as some parameter $n$, usually the sample size, increases indefinitely. For each fixed $t$, assume that $g_2(t)$ and its partial derivatives are $O(n)$ and that $g_1(t)$ is $O(1)$. Now consider a real-valued variable $\rho(\hat{\eta})$, where the function $\rho$ has continuous gradient that is nonzero at $t_{max}$.

To calculate the marginal tail probability $P(\rho(\hat{\eta}) \leq \rho_0)$, let $t_{max|\rho_0}$ be the value of $t$ that maximizes $g_2(t)$ subject to the constraint $\rho(t) = \rho_0$. Moreover, let $\rho_{max} =$

$\rho(t_{max})$, so that $\rho(\hat{\eta}) - \rho_{max}$ is $O_p(n^{-\frac{1}{2}})$ and $t_{max|\rho_{max}} = t_{max}$.

Consider the function

$$r(\rho_0) = sign(\rho_0 - \rho_{max}) \{2[g_2(t_{max}) - g_2(t_{max|\rho_0})]\}^{\frac{1}{2}},$$

which is assumed to be monotonic increasing.

An approximation to the distribution function of $\rho(\hat{\eta})$ based on normal approximations to the distribution of $R = r(\rho(\hat{\eta}))$ is as follows, provided that $\rho_0 - \rho_{max}$ is $O(n^{-\frac{1}{2}})$,

$$\begin{aligned} P(\rho(\hat{\eta}) \leq \rho_0) &= P(R \leq r_0) \\ &= \Phi(r_0) + O(n^{-\frac{1}{2}}), \end{aligned}$$

where $r_0 = r(\rho_0)$ and $\Phi$ is the standard normal distribution function.

DiCiccio and Martin proposed an adjustment to the approximation, which improves the error to order $O(n^{-\frac{3}{2}})$. Details can be found in their paper (1991). In addition, they showed that the approximation applies even if the joint density of $\hat{\eta}$ is replaced by an approximation such that

$$g_{\hat{\eta}}(t) = c \exp\{g_2(t)\} \{1 + O(n^{-\frac{3}{2}})\}$$

when $t - t_{max}$ is $O(n^{-\frac{1}{2}})$, where $c$ is a normalizing constant such that $c \exp\{g_2(t)\}$ integrates to $1 + O(n^{-\frac{3}{2}})$.

Therefore, given the joint density approximation by Field (1982), we may apply this marginal tail area approximation to obtain the required probability. In Chapter 3, we will elaborate this idea and establish a connection to our approximation.

## 2.7 Discussion

In this chapter, we have introduced two models and two estimators for discussion. The models and estimators will be seen again in our numerical examples. We have then presented several asymptotic results for the estimators, and the work on approximations by Field (1982), Tingley and Field (1990), and DiCiccio and Martin (1991), We now compare briefly the different techniques.

When the asymptotic distribution of an estimator is available, it is possibly the simplest approximation for use. In some simple applications, for example the arithmetic mean, the approximation indeed gives very reasonable results. However, the asymptotic results can be very disappointing in more complicated situations. We will give some examples in Chapter 4. In those cases, the approximation could be improved, for example, by some mean and variance adjustments, provided that the moments can be obtained. In general, the normal approximation works very well around the expected value of the estimator, and the rate of convergence of the estimator is of order $O_p(n^{\frac{1}{2}})$.

On the other hand, the joint density approximation developed by Field has been shown in many situations to give very accurate results. The major obstacle in applying this technique is its computational requirements. At each point where the joint density is needed, a system of $p$ non-linear equations must be solved, and a multiple integration is needed to obtain the required marginal distribution. While this is still manageable for low dimensional problems, it becomes impractical when the dimension is moderate or high.

The technique developed by DiCiccio and Martin is not by itself a solution to our problem. However, when the joint density function or a good approximation of

it is available, the technique becomes a useful device to avoid the multi-dimensional numerical integration. Daniels and Young (1991) found that a direct application of Laplace's method in an integration could be unacceptably inaccurate, but this tail area approximation, while it has applied the Laplace method, gives very accurate results. Therefore, a good co-operation of this technique with the joint density approximation by Field may lead to a simple and accurate result.

The linear approximation derived by Tingley and Field is clearly a nice result. It allows us to work directly on our marginal distribution problem. Numerical results show that in many cases it improves the asymptotic approximations. However, its distribution tends to be more 'normal' than the true distribution. This may be related to its nature as a mean of independent functions. Generally, the approximation provides very good approximation around the expected value of an estimator and becomes inadequate in the tails.

To summarize, the normal approximation is simple to use and works well around the expected value. A linear approximation improves it but is still inadequate in the tails. The joint density approximation suggests that we work on the density of individual points but is too computationally demanding. Combining all these remarks is exactly the idea of our approach which we are going to present next.

# Chapter 3

# Approximation for marginal densities

## 3.1 Overview

In this chapter we derive an approximation for the marginal density of a component in a general multivariate $M$-estimator. The result is generalized to a real-valued function of the estimator. Our approach is partly motivated by the results in Field (1982), and Tingley and Field (1990) (see Chapter 2). We now explain the background relationship among the three procedures. The technical connection will be clear in the development of our approximation.

Our approach originated in Field (1982) where a very accurate approximation for the joint density function of a multivariate $M$-estimator $\hat{\eta}$ is given. Although one may integrate the density approximation numerically to obtain the required marginal density, the substantial computational requirement makes it impractical even for a small dimensional problem. Tingley and Field (1990) used results in Field (1982) and derived a linear approximation $\overline{G}$ to a real-valued function of the estimator, say $\rho(\hat{\eta})$.

The problem is then reduced to one dimension and the computation becomes feasible. In spite of its simplicity, we will see in Chapter 4 that this single linear approximation may not provide a satisfactory approximation for the parts of the tail distribution of practical interest. Nevertheless, numerical results show that the distributions of $\rho(\hat{\eta})$ and $\overline{G}$ agree well near the expected value of $\rho(\hat{\eta})$. This motivates us to use $\overline{G}$ only for approximating the marginal density at the expected value rather than for the whole distribution of the function. We summarize our procedure as follows.

Consider the marginal density of $\rho(\hat{\eta})$ at $\rho_0$ under the joint density $f$.

Step 1: An exponential tilt is applied to $f$ such that under the joint conjugate density $h$, $\hat{\eta}$ is centered at $t_0$ in expectation and $\rho(t_0) = \rho_0$ for some chosen $t_0$.

Step 2: A linear function $G = \rho_0 + \overline{G}$ of the score function $\Psi$ is used to approximate $\rho(\hat{\eta})$ and give the marginal density of $\rho(\hat{\eta})$ at $\rho_0$ under $h$.

Step 3: The approximation of the marginal density under $h$ is transformed to an approximation under $f$.

In the next section, we define some general notation and state the regularity assumptions for our approximation. The main result is derived in Section 3.3. We discuss the errors of the approximation in Section 3.4, and propose some finite sample adjustments in Section 3.5. Finally in Section 3.6 we summarize the result and compare it with the techniques which were discussed in Chapter 2. In particular, we establish a technical connection between our approach and the tail probability approximation by DiCiccio and Martin (1991).

## 3.2 General notation and regularity conditions

In this section we define some basic notation for our approximation and state the regularity assumptions for the development. We begin with a brief review of the general problem.

Consider an independent random sample $Y = (Y_1, \ldots, Y_n)$, where $Y_l$, $l = 1, \ldots, n$, is $m$-dimensional and has a density function $f_l(y_l)$ that is parameterized by $\eta = (\eta_1, \ldots, \eta_p) \in \Omega \subset \Re^p$. Denote the joint density function of $Y$ by $f(y)$, so

$$f(y) = \prod_{l=1}^{n} f_l(y_l).$$

Let $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_p)$ be a multivariate $M$-estimator of $\eta$, that is, $\hat{\eta}$ solves the $p$-dimensional system

$$\frac{1}{n} \sum_{l=1}^{n} \Psi_l(Y_l, \hat{\eta}) = 0, \tag{3.1}$$

where $\Psi_l = (\Psi_{1l}, \ldots, \Psi_{pl})$, $l = 1, \ldots, n$. Note that in our location-scale model, $\Psi_l$'s are the same for all $l$, and in our multiple regression model, $\Psi_l$ depends on $l$ through the $l^{th}$ row of the design matrix.

The problem is to find an approximation for a marginal density of $\hat{\eta}$. We focus in this chapter on the derivation of an approximation for a real-valued function $\rho$ of $\hat{\eta}$ under $\eta = \eta_0$. The result includes $\rho(\hat{\eta}) = \hat{\eta}_k$ as a special case. In Chapter 5, we will extend the technique to approximate the joint density of a real-valued vector $\rho(\hat{\eta}) = (\rho_1(\hat{\eta}), \ldots, \rho_k(\hat{\eta}))$, $k \leq p$.

Define $h_{l,t_0}$ to be a conjugate density of $f_l$, that is,

$$h_{l,t_0}(y_l) = c_l(t_0)\, exp\left\{\sum_{j=1}^{p} \alpha_j \Psi_{jl}(y_l, t_0)\right\} f_l(y_l), \quad l = 1, \ldots, n,$$

where

$$c_l^{-1}(t_0) = \int_{y_l} exp\left\{\sum_{j=1}^{p} \alpha_j \Psi_{jl}(y_l, t_0)\right\} f_l(y_l)\, dy_l,$$

and $\alpha = (\alpha_1, \ldots, \alpha_p)$ is chosen so that for some fixed $t_0 = (t_{01}, \ldots, t_{0p})$,

$$E_{h_{t_0}}\left[\frac{1}{n}\sum_{l=1}^{n} \Psi_l(Y_l, t_0)\right] = 0, \tag{3.2}$$

where $h_{t_0}$ is the joint conjugate density function of $Y$ given by

$$h_{t_0}(y) = \prod_{l=1}^{n} h_{l,t_0}(y_l).$$

Note that the dependence on $t_0$ will be suppressed on both $h_{l,t_0}$ and $h_{t_0}$. The choice of $t_0$ is crucial to our approximation and will be discussed in the next section.

Let

$$\Psi_l^{(j_1 \ldots j_v)}(y_l, \eta) = \frac{\partial^v}{\partial \eta_{j_1} \ldots \partial \eta_{j_v}} \Psi_l(y_l, \eta)$$

and

$$\rho^{(j_1 \ldots j_v)}(\eta) = \frac{\partial^v}{\partial \eta_{j_1} \ldots \partial \eta_{j_v}} \rho(\eta)$$

for $1 \le j_1, \ldots, j_v \le p$, $l = 1, \ldots, n$, and

$$S = (S_1, \ldots, S_p) = \left\{\sum_{l=1}^{n} \Psi_{jl}(Y_l, t_0)\right\}_{j=1,\ldots,p}.$$

By convention, we define

$$\Psi_l^{(0)}(y_l, \eta) = \Psi_l(y_l, \eta), \quad \Psi_l^{(j_1 \ldots j_v)}(y_l, t_0) = \Psi_l^{(j_1 \ldots j_v)}(y_l, \eta)\Big|_{\eta = t_0},$$

$$\rho^{(0)}(\eta) = \rho(\eta), \quad \rho^{(j_1 \ldots j_v)}(t_0) = \rho^{(j_1 \ldots j_v)}(\eta)\Big|_{\eta = t_0}.$$

We now make eight regularity assumptions about the functions $\Psi$ and $\rho$ for our approximation. The assumptions are similar to those in Tingley and Field (1990)

with some minor changes and some adjustments in notation to accommodate our non-identical $Y_l$'s. A more general form and justifications of these assumptions can be found in Field and Ronchetti (1990).

A1 The system (3.1) has a unique solution $\hat{\eta}$.

A2 The system (3.2) has a unique solution $\alpha$ for each $t_0 \in \Omega$.

A3 The joint density of $(S, \hat{\eta})$ exists and has Fourier transforms which are absolutely integrable both under $f$ and $h$.

A4 For the $m$-dimensional $Y_l$'s, there is an open subset $U \subset \Re^m$ such that for each $\eta \in \Omega$ and $l = 1, \ldots, n$,

(a) $\int_U f_l(y_l)\, dy_l = 1$,

(b) The derivatives

$$\Psi_l^{(j_1)}(y_l, \eta), \quad \Psi_l^{(j_1 j_2)}(y_l, \eta), \quad \Psi_l^{(j_1 j_2 j_3)}(y_l, \eta)$$

exist for $1 \le j_1, j_2, j_3 \le p$.

A5 For each compact $\mathcal{K} \subset \Omega$ and $l = 1, \ldots, n$,

(a) for $0 \le j_1, j_2 \le p$,

$$\sup_{t_0 \in \mathcal{K}} E_{h|t_0}\left[\left(\Psi_l^{(j_1 j_2)}(Y_l, t_0)\right)^4\right] < \infty,$$

(b) there exists a $\delta > 0$ such that for $1 \le j_1, j_2, j_3 \le p$,

$$\sup_{t_0 \in \mathcal{K}} E_{h|t_0}\left[\max_{|\eta - t_0| \le \delta} |\Psi_l^{(j_1 j_2 j_3)}(Y_l, \eta)|^3\right] < \infty.$$

A6 For each $\eta \in \Omega$, the matrices

$$A(\eta) = \left\{ E_{h|\eta} \left[ \frac{1}{n} \sum_{l=1}^{n} \Psi_{j_1 l}^{(j_2)}(Y_l, \eta) \right] \right\}_{1 \le j_1, j_2 \le p}$$

and

$$C(\eta) = E_{h|\eta} \left[ \left( \frac{1}{n} \sum_{l=1}^{n} \Psi_l(Y_l, \eta) \right) \left( \frac{1}{n} \sum_{l=1}^{n} \Psi_l(Y_l, \eta) \right)^T \right]$$

are non-singular.

A7 The functions $A(\eta)$ and

$$E_{h|\eta} \left[ \left( \frac{1}{n} \sum_{l=1}^{n} \Psi_l^{(j_1 j_2)}(Y_l, \eta) \right) \left( \frac{1}{n} \sum_{l=1}^{n} \Psi_l^{(i_1 i_2)}(Y_l, \eta) \right)^T \right],$$

$0 \le j_1, j_2, i_1, i_2 \le p,\ j_1 + j_2 \ge 1,\ i_1 + i_2 \ge 1$, are continuous on $\Omega$.

A8 For each compact $\mathcal{K} \subset \Omega$,

(a) for $0 \le j_1, j_2, j_3 \le p$,

$$\sup_{\eta \in \mathcal{K}} |\rho^{(j_1 j_2 j_3)}(\eta)| < \infty,$$

(b) for each $t_0 \in \mathcal{K}$, there exists a $\delta > 0$ such that

$$\inf_{|\eta - t_0| < \delta} var_{h|t_0} \left[ \Psi_l^T(Y_l, \eta) \left( A^{-1}(\eta) \right)^T \frac{\partial \rho(\eta)}{\partial \eta} \right] > 0, \quad l = 1, \ldots, n.$$

## 3.3 Derivation of the approximation

In this section we derive our main result which is an approximation for the marginal density function of a real-valued function $\rho$ of a multi-dimensional $M$-estimator $\hat{\eta}$. We follow the three-step procedure as outlined in Section 3.1.

Assume that a centered conjugate density $h$ is chosen under the requirement in Step 1. Note that the $t_0$ in the condition is not specified yet. The choice of $t_0$ is an important key to our approximation and will be discussed later in the development. The next step is to derive linear approximations for $\hat{\eta}$ and $\rho(\hat{\eta})$. The construction parallels that in Field (1982) and Tingley and Field (1990).

For $1 \leq j, j_1, j_2 \leq p$ and $l = 1, \ldots, n$, define

$$\Psi_{jl} = \Psi_{jl}(Y_l, t_0), \quad \mu_{jl} = E_h[\Psi_{jl}], \quad \overline{\Psi}_j = \frac{1}{n}\sum_{l=1}^{n} \Psi_{jl}, \quad \mu_j = E_h[\overline{\Psi}_j],$$

$$\Psi_{jl}^{(j_1)} = \Psi_{jl}^{(j_1)}(Y_l, t_0), \quad \mu_{jl}^{(j_1)} = E_h[\Psi_{jl}^{(j_1)}], \quad \overline{\Psi}_j^{(j_1)} = \frac{1}{n}\sum_{l=1}^{n} \Psi_{jl}^{(j_1)}, \quad \mu_j^{(j_1)} = E_h[\overline{\Psi}_j^{(j_1)}],$$

$$\Psi_{jl}^{(j_1 j_2)} = \Psi_{jl}^{(j_1 j_2)}(Y_l, t_0), \quad \mu_{jl}^{(j_1 j_2)} = E_h[\Psi_{jl}^{(j_1 j_2)}],$$

$$\overline{\Psi}_j^{(j_1 j_2)} = \frac{1}{n}\sum_{l=1}^{n} \Psi_{jl}^{(j_1 j_2)}, \quad \mu_j^{(j_1 j_2)} = E_h[\overline{\Psi}_j^{(j_1 j_2)}].$$

Recall from the condition (3.2) that for a given $t_0$, $\alpha$ is chosen such that $\mu_j = 0$ for $j = 1, \ldots, p$. Let

$$\overline{\Psi} = (\overline{\Psi}_1, \ldots, \overline{\Psi}_p),$$

$$\overline{\Psi}^{(j)} = (\overline{\Psi}_1^{(j)}, \ldots, \overline{\Psi}_p^{(j)}), \quad \mu^{(j)} = (\mu_1^{(j)}, \ldots, \mu_p^{(j)}),$$

$$\overline{\Psi}^{(j_1 j_2)} = (\overline{\Psi}_1^{(j_1 j_2)}, \ldots, \overline{\Psi}_p^{(j_1 j_2)}), \quad \mu^{(j_1 j_2)} = (\mu_1^{(j_1 j_2)}, \ldots, \mu_p^{(j_1 j_2)}),$$

$$\overline{\Psi}^{*} = (\overline{\Psi}, \overline{\Psi}^{(1)}, \ldots, \overline{\Psi}^{(p)}, \overline{\Psi}^{(11)}, \ldots, \overline{\Psi}^{(pp)}),$$

$$\mu^* = E_h[\overline{\Psi}^*] = (0, \mu^{(1)}, \ldots, \mu^{(p)}, \mu^{(11)}, \ldots, \mu^{(pp)}).$$

A two-term Taylor series expansion of (3.1) about $t_0$ is given by

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{l=1}^{n} \Psi_l(Y_l, \hat{\eta}) \\ &= \overline{\Psi} + \sum_{j=1}^{p}(\hat{\eta} - t_0)_j \overline{\Psi}^{(j)} + \frac{1}{2}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}(\hat{\eta} - t_0)_{j_1}(\hat{\eta} - t_0)_{j_2}\overline{\Psi}^{(j_1 j_2)} + O_p|\hat{\eta} - t_0|^3. \end{aligned}$$

The approximation we obtain is actually for $\rho(T)$, where $T = (T_1, \ldots, T_p)$ is the solution of

$$\overline{\Psi} + \sum_{j=1}^{p}(T - t_0)_j \overline{\Psi}^{(j)} + \frac{1}{2}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}(T - t_0)_{j_1}(T - t_0)_{j_2}\overline{\Psi}^{(j_1 j_2)} = 0,$$

rather than for $\rho(\hat{\eta})$.

At any fixed point $Y = y = (y_1, \ldots, y_n)$, let $\psi_{jl} = \Psi(y_l, t_0)$ for $j = 1, \ldots, p$, $l = 1, \ldots, n$, and so on. Consider the system of equations

$$q(\overline{\psi}^*, t) = \overline{\psi} + \sum_{j=1}^{p}(t - t_0)_j \overline{\psi}^{(j)} + \frac{1}{2}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}(t - t_0)_{j_1}(t - t_0)_{j_2}\overline{\psi}^{(j_1 j_2)}.$$

Now, $q$ maps $\Re^{p^*+p}$ into $\Re^p$ where $p^* = p + p^2 + p^3$. Since $A(t_0)$ is non-singular by assumption A6 and

$$q(\mu^*, t_0) = 0 + \sum_{j=1}^{p}(t_0 - t_0)_j \mu^{(j)} + \frac{1}{2}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}(t_0 - t_0)_{j_1}(t_0 - t_0)_{j_2}\mu^{(j_1 j_2)} = 0,$$

the implicit function theorem can be applied to prove the existence of a unique differentiable function $H(\overline{\psi}^*)$, $H : \Re^{p^*} \to \Re^p$, such that $q(\overline{\psi}^*, H(\overline{\psi}^*)) = 0$ for $\overline{\psi}^*$ in a neighbourhood of $\mu^*$ and $H(\overline{\psi}^*)$ in a neighbourhood of $t_0$. It follows that $T = H(\overline{\Psi}^*)$ and

$$q(\overline{\Psi}^*, H(\overline{\Psi}^*)) = \overline{\Psi} + \sum_{j=1}^{p}(H(\overline{\Psi}^*) - t_0)_j \overline{\Psi}^{(j)} +$$

$$\frac{1}{2}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}(H(\overline{\Psi}^*)-t_0)_{j_1}(H(\overline{\Psi}^*)-t_0)_{j_2}\overline{\Psi}^{(j_1j_2)}$$

$$= 0. \tag{3.3}$$

We can now give a linear approximation for $T$.

**Lemma 3.1**

$$T_k = t_{0k} + \sum_{j=1}^{p}\overline{\Psi}_j B_{kj} + O_p|\overline{\Psi}^* - \mu^*|^2, \quad k = 1,\ldots,p,$$

*where* $B = \{B_{kj}\}_{1\le k,j\le p} = -A^{-1}(t_0)$ *and the first term in the error is given by*

$$e_{Tk} = \sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{j_3=1}^{p}\overline{\Psi}_{j_1}(\overline{\Psi}_{j_2}^{(j_3)} - \mu_{j_2}^{(j_3)})B_{kj_2}B_{j_3j_1} +$$

$$\frac{1}{2}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\overline{\Psi}_{j_1}\overline{\Psi}_{j_2}\sum_{j_3=1}^{p}B_{kj_3}\sum_{j_4=1}^{p}\sum_{j_5=1}^{p}B_{j_4j_2}\mu_{j_3}^{(j_4j_5)}B_{j_5j_1}.$$

**Proof** Expanding $H_k(\overline{\Psi}^*)$ in a Taylor series expansion about $\mu^*$ yields

$$H_k(\overline{\Psi}^*)$$

$$= H_k(\mu^*) + \sum_{j=1}^{p^*}(\overline{\Psi}^* - \mu^*)_j\frac{\partial H_k(\mu^*)}{\partial\overline{\Psi}_j^*} +$$

$$\frac{1}{2}\sum_{j_1=1}^{p^*}\sum_{j_2=1}^{p^*}(\overline{\Psi}^* - \mu^*)_{j_1}(\overline{\Psi}^* - \mu^*)_{j_2}\frac{\partial^2 H_k(\mu^*)}{\partial\overline{\Psi}_{j_1}^*\partial\overline{\Psi}_{j_2}^*} + O_p|\overline{\Psi}^* - \mu^*|^3$$

$$= H_k(\mu^*) + \sum_{j=1}^{p}\overline{\Psi}_j\frac{\partial H_k(\mu^*)}{\partial\overline{\Psi}_j} + \sum_{j_1=1}^{p}\sum_{j_2=1}^{p}(\overline{\Psi}_{j_1}^{(j_2)} - \mu_{j_1}^{(j_2)})\frac{\partial H_k(\mu^*)}{\partial\overline{\Psi}_{j_1}^{(j_2)}} +$$

$$\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{j_3=1}^{p}(\overline{\Psi}_{j_1}^{(j_2j_3)} - \mu_{j_1}^{(j_2j_3)})\frac{\partial H_k(\mu^*)}{\partial\overline{\Psi}_{j_1}^{(j_2j_3)}} + \frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{p}\overline{\Psi}_j\overline{\Psi}_i\frac{\partial^2 H_k(\mu^*)}{\partial\overline{\Psi}_j\partial\overline{\Psi}_i} +$$

$$\frac{1}{2}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{i_1=1}^{p}\sum_{i_2=1}^{p}(\overline{\Psi}_{j_1}^{(j_2)} - \mu_{j_1}^{(j_2)})(\overline{\Psi}_{i_1}^{(i_2)} - \mu_{i_1}^{(i_2)})\frac{\partial^2 H_k(\mu^*)}{\partial\overline{\Psi}_{j_1}^{(j_2)}\partial\overline{\Psi}_{i_1}^{(i_2)}} +$$

$$\frac{1}{2}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{j_3=1}^{p}\sum_{i_1=1}^{p}\sum_{i_2=1}^{p}\sum_{i_3=1}^{p}(\overline{\Psi}_{j_1}^{(j_2j_3)}-\mu_{j_1}^{(j_2j_3)})(\overline{\Psi}_{i_1}^{(i_2i_3)}-\mu_{i_1}^{(i_2i_3)})\frac{\partial^2 H_k(\mu^*)}{\partial\overline{\Psi}_{j_1}^{(j_2j_3)}\partial\overline{\Psi}_{i_1}^{(i_2i_3)}} \quad +$$

$$\sum_{j=1}^{p}\sum_{i_1=1}^{p}\sum_{i_2=1}^{p}\overline{\Psi}_j(\overline{\Psi}_{i_1}^{(i_2)}-\mu_{i_1}^{(i_2)})\frac{\partial^2 H_k(\mu^*)}{\partial\overline{\Psi}_j\partial\overline{\Psi}_{i_1}^{(i_2)}} \quad +$$

$$\sum_{j=1}^{p}\sum_{i_1=1}^{p}\sum_{i_2=1}^{p}\sum_{i_3=1}^{p}(\overline{\Psi}_j-\mu_j)(\overline{\Psi}_{i_1}^{(i_2i_3)}-\mu_{i_1}^{(i_2i_3)})\frac{\partial^2 H_k(\mu^*)}{\partial\overline{\Psi}_j\partial\overline{\Psi}_{i_1}^{(i_2i_3)}} \quad +$$

$$\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{i_1=1}^{p}\sum_{i_2=1}^{p}\sum_{i_3=1}^{p}(\overline{\Psi}_{j_1}^{(j_2)}-\mu_{j_1}^{(j_2)})(\overline{\Psi}_{i_1}^{(i_2i_3)}-\mu_{i_1}^{(i_2i_3)})\frac{\partial^2 H_k(\mu^*)}{\partial\overline{\Psi}_{j_1}^{(j_2)}\partial\overline{\Psi}_{i_1}^{(i_2i_3)}} \quad +$$

$$O_p|\overline{\Psi}^*-\mu^*|^3,$$

where

$$\frac{\partial H_k(\mu^*)}{\partial\overline{\Psi}_j^*}=\left.\frac{\partial H_k(\overline{\Psi}^*)}{\partial\overline{\Psi}_j^*}\right|_{\overline{\Psi}^*=\mu^*}$$

and similar interpretations for the other partial derivatives. Recall from (3.3) that

$$q_r \;=\; \overline{\Psi}_r+\sum_{j=1}^{p}(H-t_0)_j\overline{\Psi}_r^{(j)}+\frac{1}{2}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}(H-t_0)_{j_1}(H-t_0)_{j_2}\overline{\Psi}_r^{(j_1j_2)} \;=\; 0,$$

$r=1,\ldots,p$. We evaluate the partial derivatives as follows. For $1\le j\le p$,

$$\frac{\partial q_r}{\partial\overline{\Psi}_j}+\sum_{k=1}^{p}\frac{\partial q_r}{\partial H_k}\frac{\partial H_k}{\partial\overline{\Psi}_j}=0$$

implies that

$$I_{\{r=j\}}+\sum_{k=1}^{p}\left\{\overline{\Psi}_r^{(k)}+\sum_{s=1}^{p}(H-t_0)_s\overline{\Psi}_r^{(sk)}\right\}\frac{\partial H_k}{\partial\overline{\Psi}_j}=0, \tag{3.4}$$

where $I_{\{r=j\}}=1$ if $r=j$ and 0 otherwise. At $(\overline{\Psi}^*,H)=(\mu^*,t_0)$ the system (3.4) simplifies to

$$\{I_{\{r=j\}}\}_{r=1,\ldots,p}+\{\mu_r^{(k)}\}_{1\le r,k\le p}\frac{\partial H(\mu^*)}{\partial\overline{\Psi}_j}=0,\quad 1\le j\le p.$$

Recall that $A = \{\mu_r^{(k)}\}_{1\le r,k\le p}$ and define $B = -A^{-1}$. Solving the last system for the partial derivatives we obtain

$$\frac{\partial H_k(\mu^*)}{\partial \overline{\Psi}_j} = \{-A^{-1}\}_{kj} = B_{kj}.$$

In a similar manner, it can be shown that the other first order partial derivatives vanish. Now, consider the second order partial derivatives. For $1 \le j_1 \le p$, differentiating the system (3.4) with respect to $\overline{\Psi}_{j_2}$, $j_2 = 1, \ldots, p$, yields

$$\sum_{k=1}^{p}\left\{\sum_{s=1}^{p}\frac{\partial H_s}{\partial \overline{\Psi}_{j_2}}\overline{\Psi}_r^{(sk)}\frac{\partial H_k}{\partial \overline{\Psi}_{j_1}} + \left\{\overline{\Psi}_r^{(k)} + \sum_{s=1}^{p}(H-t_0)_s\overline{\Psi}_r^{(sk)}\right\}\frac{\partial^2 H_k}{\partial \overline{\Psi}_{j_1}\partial \overline{\Psi}_{j_2}}\right\} = 0.$$

At $(\overline{\Psi}^*, H) = (\mu^*, t_0)$ the above system simplifies to

$$\sum_{k=1}^{p}\left\{\sum_{s=1}^{p}B_{sj_2}\,\mu_r^{(sk)}B_{kj_1} + \mu_r^{(k)}\frac{\partial^2 H_k(\mu^*)}{\partial \overline{\Psi}_{j_1}\partial \overline{\Psi}_{j_2}}\right\} = 0$$

or equivalently,

$$\{\mu_{j_3}^{(k)}\}_{1\le j_3,k\le p}\left\{\frac{\partial^2 H_k(\mu^*)}{\partial \overline{\Psi}_{j_1}\partial \overline{\Psi}_{j_2}}\right\}_{k=1,\ldots,p} = -\left\{\sum_{j_5=1}^{p}\sum_{j_4=1}^{p}B_{j_4j_2}\,\mu_{j_3}^{(j_4j_5)}B_{j_5j_1}\right\}_{j_3=1,\ldots,p}.$$

Solving the last system for the partial derivatives gives

$$\frac{\partial^2 H_k(\mu^*)}{\partial \overline{\Psi}_{j_1}\partial \overline{\Psi}_{j_2}} = \sum_{j_3=1}^{p}B_{kj_3}\sum_{j_4=1}^{p}\sum_{j_5=1}^{p}B_{j_4j_2}\,\mu_{j_3}^{(j_4j_5)}B_{j_5j_1}.$$

Similarly, differentiating the system (3.4) with respect to $\overline{\Psi}_{j_2}^{(j_3)}$, $1 \le j_2, j_3 \le p$, yields

$$I_{\{r=j_2\}}\frac{\partial H_{j_3}}{\partial \overline{\Psi}_{j_1}} +$$

$$\sum_{k=1}^{p}\left\{\sum_{s=1}^{p}\frac{\partial H_s}{\partial \overline{\Psi}_{j_2}^{(j_3)}}\overline{\Psi}_r^{(sk)}\frac{\partial H_k}{\partial \overline{\Psi}_{j_1}} + \left\{\overline{\Psi}_r^{(k)} + \sum_{s=1}^{p}(H-t_0)_s\overline{\Psi}_r^{(sk)}\right\}\frac{\partial^2 H_k}{\partial \overline{\Psi}_{j_1}\partial \overline{\Psi}_{j_2}^{(j_3)}}\right\} = 0.$$

At $(\overline{\Psi}^*, H) = (\mu^*, t_0)$ the last system simplifies to

$$\{I_{\{r=j_2\}}B_{j_3j_1}\}_{r=1,\ldots,p} + \{\mu_r^{(k)}\}_{1\le r,k\le p}\frac{\partial^2 H(\mu^*)}{\partial\overline{\Psi}_{j_1}\partial\overline{\Psi}_{j_2}^{(j_3)}} = 0.$$

Solving the last system for the partial derivatives gives

$$\frac{\partial^2 H_k(\mu^*)}{\partial\overline{\Psi}_{j_1}\partial\overline{\Psi}_{j_2}^{(j_3)}} = B_{kj_2}B_{j_3j_1}.$$

In a similar manner, it can be shown that all the other second order partial derivatives vanish. The result follows.

□

Note that the error term $e_T$ corrects some typographical errors in equation (5) in Field (1982). Now, consider the one-term Taylor series expansion of $\rho(T)$ about $t_0$

$$\rho(T) = \rho(t_0) + (T-t_0)^T\{\mu'^{k)}(t_0)\}_{k=1,\ldots,p} + O_p|T-t_0|^2. \tag{3.5}$$

Applying Lemma 3.1 to approximate the difference $T - t_0$ in (3.5) we obtain an approximation of $\rho(T)$ as

$$\begin{aligned}\rho(T) &\approx \rho(t_0) + \left\{\sum_{j=1}^{p}\overline{\Psi}_j B_{kj}\right\}_{k=1,\ldots,p}^{T}\{\rho^{(k)}(t_0)\}_{k=1,\ldots,p} \\ &= \rho(t_0) + \frac{1}{n}\sum_{l=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{p}\rho^{(k)}(t_0)B_{kj}\Psi_{jl}(Y_l,t_0).\end{aligned} \tag{3.6}$$

We define $G$ to be the right hand side of (3.6). A one-term Edgeworth approximation to the marginal density of $G$ at $\rho_0$ under $h$ is given by

$$g_{G|h}(\rho_0) = \sqrt{\frac{1}{2\pi\sigma^2_{G|h}}}, \tag{3.7}$$

where $\sigma^2_{G|h}$ is the variance of $G$ under $h$, and the error is of order $O(n^{-\frac{1}{2}})$ (Esseen, 1945, page 44). In order to improve the normal approximation, we center $\rho(T)$ at $\rho_0$

and therefore choose $\alpha$ such that

$$E_h\left[\frac{1}{n}\sum_{l=1}^{n}\Psi_l(Y_l,t_0)\right]=0 \quad \text{and} \quad \rho(t_0)=\rho_0. \tag{3.8}$$

It follows from the expansion (3.5) and Lemma 3.1 that

$$\rho(T)-\rho_0=\left\{\sum_{j=1}^{p}\overline{\Psi}_j B_{kj}+O_p|\overline{\Psi}^*-\mu^*|^2\right\}_{k=1,\ldots,p}^{T}\{\rho^{(k)}(t_0)\}_{k=1,\ldots,p}+O_p|T-t_0|^2.$$

Taking expectations on both sides under $h$, the sum in the right hand side expression vanishes, so that we have $E_h[\rho(T)]=\rho_0$ up to the first order. It remains to choose $t_0$ and to carry out Step 3 in the approximating procedure. The following centering lemma is required.

**Lemma 3.2** *The marginal density of $\rho(\hat{\eta})$ at $\rho_0$ under $f$ and that under $h$ are related by*

$$g_f(\rho_0)=\left\{\prod_{l=1}^{n}c_l(t_0)\right\}^{-1}E_h\left[exp\left\{-\sum_{j=1}^{p}\alpha_j S_j\right\}\middle|\,\rho(\hat{\eta})=\rho_0\right]g_h(\rho_0),$$

*where*

$$S=(S_1,\ldots,S_p)=\left\{\sum_{l=1}^{n}\Psi_{jl}(Y_l,t_0)\right\}_{j=1,\ldots,p}.$$

**Proof** Denote the joint density of $(S,\hat{\eta})$ under $f$ and that under $h$ by $g_f(s,t)$ and $g_h(s,t)$ respectively. Writing

$$S=(S_1(Y),\ldots,S_p(Y)) \quad \text{and} \quad \hat{\eta}=(\hat{\eta}_1(Y),\ldots,\hat{\eta}_p(Y)),$$

the moment generating function of $(S,\hat{\eta})$ under $f$ can be written as

$$\begin{aligned}M_f(u,v) &= \int_y exp\left\{\sum_{j=1}^{p}u_j s_j(y)+\sum_{j=1}^{p}v_j t_j(y)\right\}\prod_{l=1}^{n}f_l(y_l)\,dy\\ &= \int_y exp\left\{\sum_{j=1}^{p}u_j\sum_{l=1}^{n}\Psi_{jl}(y_l,t_0)+\sum_{j=1}^{p}v_j t_j(y)\right\}\prod_{l=1}^{n}f_l(y_l)\,dy.\end{aligned}$$

Recall that

$$h_l(y_l) = c_l(t_0)\ exp\left\{\sum_{j=1}^{p} \alpha_j \Psi_{jl}(y_l, t_0)\right\} f_l(y_l),$$

therefore by choosing $(u, v) = (\alpha + i\gamma, i\lambda)$ for some constants $\alpha$ we have

$$\begin{aligned}
& M_f(\alpha + i\gamma, i\lambda) \\
= & \int_y exp\left\{\sum_{j=1}^{p}(\alpha + i\gamma)_j \sum_{l=1}^{n} \Psi_{jl}(y_l, t_0) + \sum_{j=1}^{p} i\lambda_j t_j(y)\right\} \prod_{l=1}^{n} f_l(y_l)\ dy \\
= & \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} \int_y exp\left\{\sum_{j=1}^{p} i\gamma_j \sum_{l=1}^{n} \Psi_{jl}(y_l, t_0) + \sum_{j=1}^{p} i\lambda_j t_j(y)\right\} \prod_{l=1}^{n} h_l(y_l)\ dy \\
= & \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} M_h(i\gamma, i\lambda),
\end{aligned}$$

where $M_h(i\gamma, i\lambda)$ is the moment generating function of $(S, \hat{\eta})$ under $h$. Since both $M_f$ and $M_h$ are absolutely integrable by assumption A3, we can apply the Fourier inversion formula to obtain

$$g_f(s,t) = \frac{1}{(2\pi)^{2p}} \int_v \int_u exp\left\{-\sum_{j=1}^{p} u_j s_j - \sum_{j=1}^{p} v_j t_j\right\} M_f(u, v)\ du\ dv,$$

where the components of $u$ and $v$ are integrated along the path from $w - i\infty$ to $w + i\infty$ for some $w$. Choosing $(u, v) = (\alpha + i\gamma, i\lambda)$ yields

$$\begin{aligned}
g_f(s,t) &= \frac{1}{(2\pi)^{2p}} \int_\lambda \int_\gamma exp\left\{-\sum_{j=1}^{p}(\alpha + i\gamma)_j s_j - \sum_{j=1}^{p} i\lambda_j t_j\right\} M_f(\alpha + i\gamma, i\lambda)\ d\gamma\ d\lambda \\
&= \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} exp\left\{-\sum_{j=1}^{p} \alpha_j s_j\right\} \times \\
&\quad \frac{1}{(2\pi)^{2p}} \int_\lambda \int_\gamma exp\left\{-\sum_{j=1}^{p} i\gamma_j s_j - \sum_{j=1}^{p} i\lambda_j t_j\right\} M_h(i\gamma, i\lambda)\ d\gamma\ d\lambda \\
&= \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} exp\left\{-\sum_{j=1}^{p} \alpha_j s_j\right\} g_h(s,t).
\end{aligned}$$

Integrating both sides of the last equality along $\rho(t) = \rho_0$ gives

$$g_f(s, \rho_0) = \left\{ \prod_{l=1}^{n} c_l(t_0) \right\}^{-1} exp\left\{ -\sum_{j=1}^{p} \alpha_j s_j \right\} g_h(s, \rho_0).$$

Integrating both sides again with respect to $s$, we obtain

$$g_f(\rho_0) = \left\{ \prod_{l=1}^{n} c_l(t_0) \right\}^{-1} \int_s exp\left\{ -\sum_{j=1}^{p} \alpha_j s_j \right\} g_h(s|\rho_0)\, ds\; g_h(\rho_0)$$

and the result follows.

□

Note that we may encounter a situation in which the linear approximation for the estimator is exact. An example can be found in the next section. In that case, the joint density of $(S, \hat{\eta})$ degenerates and a slight modification is needed in the proof. To illustrate the idea, we assume that the linear approximation for $(\hat{\eta}_1, \ldots, \hat{\eta}_q)$ is exact. Define $S^* = (S_{q+1}, \ldots, S_p)$. Applying an analogous argument we obtain

$$g_f(s^*, t) = \left\{ \prod_{l=1}^{n} c_l(t_0) \right\}^{-1} exp\left\{ -\sum_{j=1}^{p} \alpha_j s_j \right\} g_h(s^*, t).$$

It follows that

$$g_f(s, t) = \left\{ \prod_{l=1}^{n} c_l(t_0) \right\}^{-1} exp\left\{ -\sum_{j=1}^{p} \alpha_j s_j \right\} g_h(s, t)$$

and we can proceed as before.

Now, combining Lemmas 3.1 and 3.2 with the Edgeworth approximation in (3.7) yields an approximation to the marginal density of $\rho(\hat{\eta})$ under $f$, that is,

$$g_f(\rho_0) \approx \left\{ \prod_{l=1}^{n} c_l(t_0) \right\}^{-1} E_h\left[ exp\left\{ -\sum_{j=1}^{p} \alpha_j S_j \right\} \middle| \rho(T) = \rho_0 \right] \sqrt{\frac{1}{2\pi\sigma^2_{G|h}}}\,.$$

However, a direct evaluation of the conditional expectation in the approximation is not very attractive. We now show that some suitable choices of $\alpha$ and $t_0$ will make the evaluation straightforward.

We need a total of $2p$ constraints to define the conjugate density $h$, that is, $p$ for the $\alpha$ and $p$ for the $t_0$. The conditions in (3.8) specify $p+1$ of them. For the relationship in Lemma 3.2, a desired situation would be when the conditional expectation

$$E_h\left[exp\left\{-\sum_{j=1}^{p}\alpha_j S_j\right\}\middle|\, \rho(\hat{\eta})=\rho_0\right]=1.$$

This is trivial if

$$\sum_{j=1}^{p}\alpha_j S_j \;\propto\; \rho(\hat{\eta})-\rho_0. \tag{3.9}$$

Using $G$ to approximate $\rho(\hat{\eta})$ and the definition of $S$, the proportionality (3.9) becomes

$$\sum_{l=1}^{n}\sum_{j=1}^{p}\alpha_j\Psi_{jl}(Y_l,t_0) \;\propto\; \sum_{l=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{p}\rho^{(k)}(t_0)B_{kj}\Psi_{jl}(Y_l,t_0),$$

which is true if

$$\alpha_{j_1}\sum_{k=1}^{p}\rho^{(k)}(t_0)B_{kj_2}=\alpha_{j_2}\sum_{k=1}^{p}\rho^{(k)}(t_0)B_{kj_1} \tag{3.10}$$

for $1\le j_1, j_2\le p$, which accounts for $p-1$ constraints.

We assume that at each point $\rho(\hat{\eta})=\rho_0$, the $2p$ constraints in (3.8) and (3.10) for choosing $\alpha$ and $t_0$ lead to a unique solution, that is, the joint conjugate density function $h$ exists and is unique. Justifications of the assumption will be given in Section 3.6. Now, putting the results together, we have the following.

**Theorem 3.1** *Let $\rho(\hat{\eta})$ be a real-valued function of a multivariate M-estimator $\hat{\eta}=(\hat{\eta}_1,\ldots,\hat{\eta}_p)$ which solves the system of equations*

$$\frac{1}{n}\sum_{l=1}^{n}\Psi_l(Y_l,\hat{\eta})=0,$$

*where $\Psi_l=(\Psi_{1l},\ldots,\Psi_{pl})$, and the $Y_l$'s are independent with densities $f_l(y_l)$. If assumptions A1 - A8 are satisfied, an approximation for the marginal density of $\rho(\hat{\eta})$*

*at $\rho_0$ under the joint density function $f = \prod_l f_l$ is given by*

$$g_p(\rho_0) = \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} \sqrt{\frac{1}{2\pi\sigma^2_{G|h}}}, \tag{3.11}$$

*where*

$$c_l^{-1}(t_0) = \int_{-\infty}^{\infty} exp\left\{\sum_{j=1}^{p} \alpha_j \Psi_{jl}(y_l, t_0)\right\} f_l(y_l)\, dy_l,$$

*$\sigma^2_{G|h}$ is the variance of*

$$G = \rho(t_0) + \frac{1}{n}\sum_{l=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{p} \rho^{(k)}(t_0) B_{kj} \Psi_{jl}(Y_l, t_0)$$

*under the joint conjugate density function $h = \prod_l h_l$,*

$$h_l(y_l) = c_l(t_0)\, exp\left\{\sum_{j=1}^{p} \alpha_j \Psi_{jl}(y_l, t_0)\right\} f_l(y_l),$$

*$\alpha$ and $t_0$ are chosen such that*

$$E_h\left[\frac{1}{n}\sum_{l=1}^{n} \Psi_l(Y_l, t_0)\right] = 0, \quad \rho(t_0) = \rho_0,$$

$$\alpha_{j_1} \sum_{k=1}^{p} \rho^{(k)}(t_0) B_{kj_2} \;=\; \alpha_{j_2} \sum_{k=1}^{p} \rho^{(k)}(t_0) B_{kj_1}, \quad 1 \le j_1, j_2 \le p,$$

*and*

$$B = -A^{-1}(t_0), \quad A(t_0) = E_h\left[\frac{1}{n}\sum_{l=1}^{n} \left.\frac{\partial \Psi_l(Y_l, \eta)}{\partial \eta^T}\right|_{\eta = t_0}\right].$$

□

In general, the approximation $g_p$ in (3.11) has to be normalized to give a total probability of one. Numerical results show that the normalization gives more accurate approximations. We define $G_p$ to be the approximation for which the density at $\rho_0$ is the normalized $g_p(\rho_0)$.

## 3.4 Errors in the approximation

We have derived $g_p$ for approximating the marginal density $g_f$ of $\rho(\hat{\eta})$. In this section, we discuss the errors which are induced in the development and give some remarks on their overall effects.

In the derivation of $g_p$, we have basically applied two approximations. The first and perhaps more important one is to use the linear approximation $G$ for the function $\hat{\rho} = \rho(\hat{\eta})$. Writing

$$\hat{\rho}_n = G_n + R_n$$

with the subscripts to emphasize the dependence to the sample size, Tingley and Field (1990) have shown that the error $R_n$ is of order $o_p(n^{-\frac{1}{2}})$ for identically distributed $Y_l$'s, and Tingley (1992) has shown that the same result holds for multiple regression problems. We need to understand how the error affects the density approximation.

Recall from Lemma 3.2 that

$$g_f(\rho_0) = \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} E_h\left[exp\left\{-\sum_{j=1}^{p}\alpha_j S_j\right\}\middle|\hat{\rho}_n = \rho_0\right] g_h(\rho_0)\,.$$

In evaluating the conditional expectation and the marginal density $g_h(\rho_0)$, we have applied $G_n$ twice.

We first consider the conditional expectation. With the conditions in Theorem 3.1 for choosing $\alpha$ and $t_0$, we can write

$$\sum_{j=1}^{p}\alpha_j S_j = d_n(G_n - \rho_0),$$

where

$$d_n = \frac{n\alpha_j}{\sum_{k=1}^{p}\rho^{(k)}(t_0)B_{kj}},\quad j = 1,\ldots,p.$$

Therefore

$$\begin{aligned} E_h\left[exp\left\{-\sum_{j=1}^{p}\alpha_j S_j\right\}\middle|\,\hat{\rho}_n=\rho_0\right] &= E_h\left[exp\{-d_n(G_n-\rho_0)\}\,|\,G_n+R_n=\rho_0\right] \\ &= E_h\left[exp\{d_nR_n\}\,|\,\hat{\rho}_n=\rho_0\right] \\ &= E_h\left[1+d_nR_n+O_p(d_n^2R_n^2)\middle|\,\hat{\rho}_n=\rho_0\right] \\ &= 1+E_h\left[o_p\left(\frac{d_n}{n^{\frac{1}{2}}}\right)\middle|\,\hat{\rho}_n=\rho_0\right]. \end{aligned}$$

Writing

$$g_h(\rho_0)=g_{G_n|h}(\rho_0)\{1+e_n\},$$

where $g_{G_n|h}$ is the density of $G_n$ at $\rho_0$ under $h$, we obtain

$$g_f(\rho_0)=\left\{\prod_{l=1}^{n}c_l(t_0)\right\}^{-1}\left\{1+E_h\left[o_p\left(\frac{d_n}{n^{\frac{1}{2}}}\right)\middle|\,\hat{\rho}_n=\rho_0\right]\right\}g_{G_n|h}(\rho_0)\{1+e_n\}.$$

Unfortunately, the convergence of $\hat{\rho}_n-G_n$ in probability generally does not give us a clue to determine the order of the expectation and $e_n$. The overall rate of error is still under investigation. However, for the limiting case with $c=\infty$ in our multiple regression model, we have the following result. (Note that in this case, the estimator $\hat{\eta}$ is the least squares estimator and $\hat{\theta}$ is a linear combination of the random observations.)

**Proposition 3.1** *For the least squares estimator $\hat{\theta}_k$, $k=1,\ldots,p-1$, of our multiple regression model, the linear approximation $G$ is exact.*

**Proof** Let

$$r_l(\eta)=\frac{Y_l-X_l^T\theta}{\sigma},\quad l=1,\ldots,n.$$

The least squares estimator corresponds to the score functions

$$\Psi_{jl}(Y_l, \eta) = r_l(\eta) X_{lj}, \quad j = 1, \ldots, p-1,$$

$$\Psi_{pl}(Y_l, \eta) = r_l^2(\eta) - \beta.$$

Therefore $t_0$ is the solution of

$$\frac{1}{n} \sum_{l=1}^{n} E_h[r_l(t_0)] X_{lj} = 0, \quad j = 1, \ldots, p-1,$$

$$\frac{1}{n} \sum_{l=1}^{n} E_h[r_l^2(t_0)] - \beta = 0.$$

Let $r_0 = (r_1(t_0), \ldots, r_n(t_0))$. By definition we have

$$A = -\frac{1}{nt_{0p}} E_h \begin{bmatrix} X^T X & X^T r_0 \\ 2r_0^T X & 2r_0^T r_0 \end{bmatrix} = -\frac{1}{nt_{0p}} \begin{bmatrix} X^T X & 0 \\ 0 & 2n\beta \end{bmatrix},$$

where the last equality follows from the definition of $t_0$, and

$$B = \begin{bmatrix} nt_{0p}(X^T X)^{-1} & 0 \\ 0 & \dfrac{t_{0p}}{2\beta} \end{bmatrix}.$$

Let $G_\theta = (G_1, \ldots, G_{p-1})$ be the approximation for $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_{p-1})$, and $t_\theta = (t_{01}, \ldots, t_{0p-1})$. For $k = 1, \ldots, p-1$, we have

$$\begin{aligned} G_k &= t_{0k} + \frac{1}{n} \sum_{l=1}^{n} \sum_{j=1}^{p} B_{kj} \Psi_{jl}(Y_l, t_0) \\ &= t_{0k} + \frac{1}{n} \sum_{l=1}^{n} \sum_{j=1}^{p-1} nt_{0p} (X^T X)^{-1}_{kj} \left( \frac{Y_l - X_l^T t_\theta}{t_{0p}} \right) X_{lj}, \end{aligned}$$

or equivalently,

$$G_\theta = t_\theta + (X^T X)^{-1} X^T (Y - X t_\theta) = (X^T X)^{-1} X^T Y$$

which is exact.

□

It follows that the proportionality (3.9) is exact under our choice of $(\alpha, t_0)$ and therefore

$$g_f(t_{0k}) = \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} g_{G|h}(t_{0k})$$

from Lemma 3.2.

We use linear approximations in the development and can expect the approximation to perform the best when $\rho(\hat{\eta})$ can be expressed as a linear combination of the score functions. The above result is a good demonstration. However, we may also want to know what happens when the estimator is not a linear combination of the score functions. We now give an example to illustrate it. Consider the least squares estimator

$$\rho(\hat{\eta}) = \hat{\sigma}^2 = \frac{\sum_{l=1}^{n}(Y_l - X_l^T\hat{\theta})^2}{n\beta}.$$

We have $\rho^{(p)}(t_0) = 2t_{0p}$ and therefore

$$\begin{aligned} G &= t_{0p}^2 + \frac{1}{n}\sum_{l=1}^{n} 2t_{0p}\,\frac{t_{0p}}{2\beta}\left\{\left(\frac{Y_l - X_l^T t_\theta}{t_{0p}}\right)^2 - \beta\right\} \\ &= \frac{\sum_{l=1}^{n}(Y_l - X_l^T t_\theta)^2}{n\beta} \end{aligned}$$

which is simply a one-term Taylor series expansion of the true function and has an error of order $o_p(n^{-\frac{1}{2}})$ as expected.

Besides using the linear function $G$ in the development of $q_p$, we use a one-term Edgeworth approximation for the density of $G$ at $\rho_0$, that is,

$$g_{G|h}(\rho_0) = \sqrt{\frac{1}{2\pi\sigma_{G|h}^2}} + error\ .$$

We know from the discussion in Section 1.4 that the error is generally of order $O(n^{-\frac{1}{2}})$. When the approximation is evaluated at the expected value, which is in our case, the error is improved to order of $O(n^{-1})$. This enables us to approximate the density accurately. For instance, applying a one-term Edgeworth approximation for the least squares estimator $\hat{\theta}_k$ yields

$$g_f(t_{0k}) = \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} \left\{\sqrt{\frac{1}{2\pi\sigma^2_{G|h}}} + O\left(\frac{1}{n}\right)\right\}.$$

Nevertheless, we need to point out that the density approximation for $g_{G|h}(\rho_0)$ is independent of the other steps in the process. Therefore if there exists a better approximation in a specific case, we can always replace the Edgeworth approximation with it.

## 3.5 Adjustments to the approximation

In our approximation the centering procedure plays a very important role. We exploit the good performance of $G$ around the center of $T$ and use an Edgeworth expansion to approximate the density at the expected value. In Lemma 3.1, $T$ is expanded about its mean $E_h[T]$ under $h$. We approximate $E_h[T]$ by $t_0$ which is correct up to the first order in the expansion, but the result should be improved if $t_0$ is replaced by a better approximation of the mean.

Recall from Lemma 3.1 that a Taylor series expansion of $T_k$ is in the form

$$T_k = t_{0k} + \sum_{j=1}^{p} \overline{\Psi}_j B_{kj} + e_{Tk} + \cdots, \quad k = 1, \ldots, p,$$

where $e_{Tk}$ is a function of $B_{j_1 j_2}$, $\mu_{j_1}^{(j_2 j_3)}$, $\overline{\Psi}_{j_1}\overline{\Psi}_{j_2}$ and $\overline{\Psi}_{j_1}(\overline{\Psi}_{j_2}^{(j_3)} - \mu_{j_2}^{(j_3)})$, $1 \le j_1, j_2, j_3 \le p$. Currently we approximate $T_k - t_{0k}$ by $\sum_j \overline{\Psi}_j B_{kj}$. A natural choice would be to replace $t_0$ by

$$t_\mu = t_0 + E_h[e_T]. \tag{3.12}$$

To implement the adjustment $E_h[e_T]$, we need to compute $B_{j_1 j_2}$ and the expected values

$$E_h[\overline{\Psi}_{j_1}^{(j_2 j_3)}], \quad E_h[\overline{\Psi}_{j_1}\overline{\Psi}_{j_2}] \text{ and } E_h[\overline{\Psi}_{j_1}(\overline{\Psi}_{j_2}^{(j_3)} - \mu_{j_2}^{(j_3)})].$$

Since the random observations are independent by assumption and the expected value of $\overline{\Psi}_{j_1}$ equals zero by the condition (3.8), it follows that

$$\begin{aligned} E_h[\overline{\Psi}_{j_1}\overline{\Psi}_{j_2}] &= \frac{1}{n^2} E_h\left[\sum_{l_1=1}^{n}(\Psi_{j_1 l_1} - \mu_{j_1 l_1}) \sum_{l_2=1}^{n}(\Psi_{j_2 l_2} - \mu_{j_2 l_2})\right] \\ &= \frac{1}{n^2}\sum_{l=1}^{n} E_h[(\Psi_{j_1 l} - \mu_{j_1 l})(\Psi_{j_2 l} - \mu_{j_2 l})] \end{aligned}$$

$$= \frac{1}{n^2}\sum_{l=1}^{n}\{E_h[\Psi_{j_1 l}\Psi_{j_2 l}] - \mu_{j_1 l}\mu_{j_2 l}\},$$

and similarly,

$$E_h[\overline{\Psi}_{j_1}(\overline{\Psi}_{j_2}^{(j_3)} - \mu_{j_2}^{(j_3)})] = \frac{1}{n^2}\sum_{l=1}^{n}\{E_h[\Psi_{j_1 l}\Psi_{j_2 l}^{(j_3)}] - \mu_{j_1 l}\mu_{j_2 l}^{(j_3)}\}.$$

Note that $B_{j_1 j_2}$, $\mu_{jl}$ and $\mu_{j_1 l}^{(j_2)}$ are required for our basic approximation, t' erefore to compute the adjustment, additional computations are needed to evaluate

$$\mu_{j_1 l}^{(j_2 j_3)},\ E_h[\Psi_{j_1 l}\Psi_{j_2 l}] \text{ and } E_h[\Psi_{j_1 l}\Psi_{j_2 l}^{(j_3)}],\ 1 \le j_1, j_2, j_3 \le p,\ l = 1, \ldots, n.$$

Recall that the conjugate density $h$ depends on $t_0$ so that a different adjustment is made for each $t_0$. This clearly increases substantially the computational requirement for the approximation.

An alternative is to use a constant adjustment so that the adjustment needs to be computed only once for all $t_0$. For this alternative, a simple choice would be

$$t_\mu = t_0 + E_f[e_T]. \tag{3.13}$$

We will see that this constant adjustment is very useful in some situations.

The above replacements are expected to give better approximations for the expectations $E_h[T]$ and $E_h[\rho(T)]$. The aim is to improve the approximation $t_\mu$ for $E_h[T]$ and also the approximation $E_h[T]$ for $E_h[\hat{\eta}]$. A desired situation is when the differences $E_h[T] - t_\mu$ and $E_h[\hat{\eta}] - E_h[T]$ are as small as possible. Since the expectations of $T$ and $\hat{\eta}$ under $h$ are not yet available for small sample problems, a sensible approach would be to reduce the errors introduced in each step of the development.

In the derivation, we applied Taylor series approximations which involve, directly or indirectly, expansions of the score functions. In general, the technique yields a very

good approximation around the point of expansion but not necessarily otherwise. To illustrate our point, consider an expansion of the Huber's score function for a simple location problem

$$\Psi_c(Y-\hat{\theta}) \approx \Psi_c(Y-t_0) + (\hat{\theta}-t_0)\left.\frac{d\Psi_c(Y-\theta)}{d\theta}\right|_{\theta=t_0} + \frac{1}{2}(\hat{\theta}-t_0)^2\left.\frac{d^2\Psi_c(Y-\theta)}{d\theta^2}\right|_{\theta=t_0},$$

where $\Psi_c(r) = max\{-c, min\{c, r\}\}$. The approximation on the right hand side is exact if $|Y-t_0| < c$ and $|Y-\hat{\theta}| < c$, and equals $c\,sign\{Y-t_0\}$ if $|Y-t_0| > c$. In the latter case the discrepancy between the approximation and the true value is

$$c\,sign\{Y-t_0\} - \Psi_c(Y-\hat{\theta})$$

which has a maximum value of $2c$. In order to obtain a better approximation for the expected value of the score function, we want to keep the expansion unchanged when it is exact and reduce the discrepancy wherever possible. With this objective in mind, we propose the following refinement for our examples. Consider $e_T = e_T(\Psi_c, \Psi_c', \Psi_c'')$. Let

$$t_\mu = t_0 + E_h[e_T(\Psi_c I_c, \Psi_c' I_c, \Psi_c'' I_c)], \tag{3.14}$$

where $I_c(x) = 1$ if $|x| < c$ and 0 otherwise. Therefore the expansion vanishes and the maximum discrepancy reduces to

$$max|0 - \Psi_c(Y-\hat{\theta})| = c$$

in the region $|Y-t_0| > c$. Note that this proposal is equivalent to the adjustment (3.12) when $c$ is taken to be infinity.

The adjustments (3.12), (3.13), and (3.14) will be implemented and compared in the numerical examples in the next chapter. Note that when $t_0$ is replaced by $t_\mu$, the

centering constraint $\rho(t_0) = \rho_0$ is changed to $\rho(t_\mu) = \rho_0$ and

$$\begin{aligned}\rho(T) - \rho_0 &\approx (T - t_\mu)\frac{\partial \rho(t_\mu)}{\partial \eta} \\ &= \frac{1}{n}\sum_{l=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{p}\Psi_{jl}(Y_l, t_0)B_{kj}(t_0)\frac{\partial \rho(t_\mu)}{\partial \eta_k}.\end{aligned}$$

The details of their computation are given in Appendix A.

Besides mean adjustment, one may also think of the possibility of reparameterization to improve the approximation. However, before considering this approach, we state an invariant property of our approximation.

**Proposition 3.2** *Let $\eta^*$ be a one-to-one reparameterization of $\eta$. Define*

$$\Psi_{jl} = \Psi_{jl}(Y_l, \eta), \quad \Psi^*_{jl} = \Psi_{jl}(Y_l, \eta(\eta^*)), \quad j = 1, \ldots, p, \quad l = 1, \ldots, n,$$

$$\rho = \rho(\eta), \quad \rho^* = \rho(\eta(\eta^*)).$$

*Denoted by $\hat{\eta}$ and $\hat{\eta}^*$ the solutions of*

$$\left\{\frac{1}{n}\sum_{l=1}^{n}\Psi_{jl} = 0\right\}_{j=1,\ldots,p} \quad \text{and} \quad \left\{\frac{1}{n}\sum_{l=1}^{n}\Psi^*_{jl} = 0\right\}_{j=1,\ldots,p}$$

*respectively. Let $\hat{\rho} = \rho(\hat{\eta})$, $\hat{\rho}^* = \rho(\eta(\hat{\eta}^*))$, and $G$ and $G^*$ be the linear approximations for $\hat{\rho}$ and $\hat{\rho}^*$ respectively. At any point $\hat{\rho}^* = \rho_0$, we have*

$$g^*_p(\rho_0) = g_p(\rho_0),$$

*where $g^*_p$ and $g_p$ are the density approximations based on $G^*$ and $G$ respectively.*

**Proof** The density of $\hat{\rho}$ at $\rho_0$ is approximated by using

$$G = \rho_0 + \overline{\Psi}^T B^T \left.\frac{\partial \rho}{\partial \eta}\right|_{\eta = t_0}$$

where the expectations are computed with the conjugate density $h$ and the $(\alpha, t_0)$ satisfies the centering conditions stated in Theorem 3.1. We need to show that the $G^*$ for $\hat{\rho}^*$ at $\rho_0$ under $h^*$ is the same as the $G$ for $\hat{\rho}$ at $\rho_0$ under $h$. Choosing $(\alpha^*, t_0^*) = (\alpha, \eta^*(t_0))$, we show that $h^*$ satisfies the centering conditions. For $j = 1, \ldots, p$ and $l = 1, \ldots, n$,

$$\Psi_{jl}^*\Big|_{\eta^*=t_0^*} = \Psi_{jl}(Y_l, \eta(\eta^*))|_{\eta(\eta^*)=t_0} = \Psi_{jl}|_{\eta=t_0}.$$

It follows that

$$E_{h^*}\left[\frac{1}{n}\sum_{l=1}^{n}\Psi_{jl}^*\Big|_{\eta^*=t_0^*}\right] = E_h\left[\frac{1}{n}\sum_{l=1}^{n}\Psi_{jl}\Big|_{\eta=t_0}\right] = 0,$$

and

$$\rho^*|_{\eta^*=t_0^*} = \rho(\eta(\eta^*))_{\eta(\eta^*)=t_0} = \rho_0.$$

For the proportionalities, we have

$$\Psi_{jl}^{*(k)} = \frac{\partial\Psi_{jl}^*}{\partial\eta_k^*} = \sum_{i=1}^{p}\frac{\partial\Psi_{jl}^*}{\partial\eta_i}\frac{\partial\eta_i}{\partial\eta_k^*} = \sum_{i=1}^{p}\frac{\partial\Psi_{jl}}{\partial\eta_i}\frac{\partial\eta_i}{\partial\eta_k^*} = \sum_{i=1}^{p}\Psi_{jl}^{(i)}\frac{\partial\eta_i}{\partial\eta_k^*}$$

which implies that

$$A^* = E\left[\frac{\partial\overline{\Psi}^*}{\partial\eta^*}\Big|_{\eta^*=t_0^*}\right] = E\left[\frac{\partial\overline{\Psi}}{\partial\eta}\frac{\partial\eta}{\partial\eta^*}\Big|_{\eta^*=t_0^*}\right] = AD$$

and

$$B^* = -A^{*-1} = D^{-1}B,$$

where

$$D = \frac{\partial\eta}{\partial\eta^*}\Big|_{\eta^*=t_0^*}.$$

Therefore

$$\alpha_{j_1} \sum_{j_3=1}^{p} \left.\frac{\partial \rho(\eta)}{\partial \eta_{j_3}}\right|_{\eta=t_0} B_{j_3 j_2} = \alpha_{j_2} \sum_{j_3=1}^{p} \left.\frac{\partial \rho(\eta)}{\partial \eta_{j_3}}\right|_{\eta=t_0} B_{j_3 j_1}$$

is equivalent to

$$\alpha_{j_1} \sum_{j_3=1}^{p} \left.\frac{\partial \rho(\eta)}{\partial \eta_{j_3}}\right|_{\eta=t_0} \sum_{k=1}^{p} \frac{\partial \eta_{j_3}}{\partial \eta_k^*} B^*_{k j_2} = \alpha_{j_2} \sum_{j_3=1}^{p} \left.\frac{\partial \rho(\eta)}{\partial \eta_{j_3}}\right|_{\eta=t_0} \sum_{k=1}^{p} \frac{\partial \eta_{j_3}}{\partial \eta_k^*} B^*_{k j_1}$$

which can be simplified to

$$\alpha_{j_1} \sum_{k=1}^{p} \left.\frac{\partial \rho(\eta(\eta^*))}{\partial \eta_k^*}\right|_{\eta^*=t_0^*} B^*_{k j_2} = \alpha_{j_2} \sum_{k=1}^{p} \left.\frac{\partial \rho(\eta(\eta^*))}{\partial \eta_k^*}\right|_{\eta^*=t_0^*} B^*_{k j_1}.$$

This implies that $(\alpha, \eta^*(t_0))$ satisfies the required centering conditions. Hence, the uniqueness of $h^*$ allows us to conclude that $h^* = h$. To complete the proof, we can write

$$G^* = \rho_0 + \overline{\Psi}^* B^{*T} \left.\frac{\partial \rho^*}{\partial \eta^*}\right|_{\eta^*=t_0^*} = \rho_0 + \overline{\Psi} B^T D^{-1^T} \left.\frac{\partial \rho^*}{\partial \eta^*}\right|_{\eta^*=t_0^*} = G$$

by realizing that

$$D^T \left.\frac{\partial \rho}{\partial \eta}\right|_{\eta=t_0} = \left\{\frac{\partial \eta}{\partial \eta^*}\right\}^T \left.\frac{\partial \rho}{\partial \eta}\right|_{\eta=t_0} = \left.\frac{\partial \rho^*}{\partial \eta^*}\right|_{\eta^*=t_0^*}.$$

The result now follows.

□

Hence, any attempt to improve the approximation by reparameterizations will not be successful. On the other hand, if a reparameterization simplifies the computation, we can always apply it without worrying about the cost in the accuracy.

## 3.6 Discussion

In this chapter, we have derived an approximation $G_p$ for the general problem as stated in Section 3.2. We have shown how the approximation is technically related to the work by Field (1982) and Tingley and Field (1990). We now give a general comparison of the three approaches.

Since both $\overline{G}$ and $G_p$ originated in the work of Field, we will not be surprised to see that they share many similarities. In fact, the initial developments of the three approaches are almost the same. The first step is to approximate an $M$-estimator $\hat{\eta}$ via a Taylor series expansion of the system where $\hat{\eta}$ is implicitly defined. After that, the three procedures are developed differently.

Tingley and Field use the first term of the expansion to construct robust confidence intervals. The objective of their work is not to approximate the marginal density of an estimator. Nevertheless, they derived a linear approximation $\overline{G}$ for a real-valued function $\rho(\hat{\eta})$ and showed that the error is of order $o_p(n^{-\frac{1}{2}})$. Their idea will be discussed further in Chapter 6.

Field established the critical link between the joint density under $f$ and that under a centered $h$. This allows him to focus on the central density approximation. His work involves the second term in the Taylor series expansion. The performance of the approximation has been shown in various applications to give very accurate results (see Field, 1982, Field and Ronchetti, 1990).

Our work is partly motivated by the link derived by Field. We found that a similar link exists between the marginal densities under $f$ and a carefully chosen $h$. We can then enjoy the good performance of a central density approximation and at the same time reduce a multi-dimensional problem to one dimension.

In terms of computational effort, the approach using $\overline{G}$ is the simplest one. The major effort goes into the evaluation of its distribution. For this purpose, Tingley and Field applied a device by Lugannani and Rice (1980) and obtained very satisfactory results. On the other hand, the work of Field requires a high dimensional integration. At each point where the joint density is computed, the expectations of some first and second order partial derivatives are needed and a system of $p$ non-linear equations must be solved. Our work does not require the expectation of any higher order partial derivative. In addition, we reduce the problem to one dimension. However, at each point where the marginal density is to be approximated, a system of $2p$ non-linear equations must be solved. Compared to the approach by Field, our approach reduces the dimension of integration but increases the size of the non-linear system. This trade off will be discussed further in Chapter 5.

To conclude this chapter, we present a connection with the probability approximation derived by DiCiccio and Martin (1991). When the joint density function of a multivariate $M$-estimator is available, we can use results of DiCiccio and Martin to obtain a marginal distribution approximation. We now establish a relationship between the two approaches. Denoting by $g_{\hat{\eta}}(t_0)$ the joint density function of $\hat{\eta}$, DiCiccio and Martin (1991) have derived an asymptotic approximation of $P(\rho(\hat{\eta}) < \rho_0)$ when

$$g_{\hat{\eta}}(t_0) \propto g_1(t_0)\ exp\{g_2(t_0)\}$$

satisfies some general conditions. In their approximation it is required to locate the maximum of $g_2(t_0)$ subject to the constraint $\rho(t_0) = \rho_0$. Taking $g_T(t_0)$ in Theorem 2.2 as an approximation of $g_{\hat{\eta}}(t_0)$, the part of the approximation corresponding to $g_2(t_0)$ would be

$$\sum_{l=1}^{n} log\left\{c_l^{-1}(t_0)\right\}.$$

Define the Lagrangian

$$L(t_0, \lambda) = \sum_{l=1}^{n} log\left\{c_l^{-1}(t_0)\right\} + \lambda(\rho(t_0) - \rho_0) = 0.$$

Differentiating $L$ with respect to $t_0$ yields

$$\begin{aligned}\frac{\partial L(t_0,\lambda)}{\partial t_0} &= \sum_{l=1}^{n} c_l(t_0) \int_{y_l} \left\{ \frac{\partial \alpha^T(t_0)}{\partial t_0}\Psi(y_l, t_0) + \frac{\partial \Psi^T(y_l, t_0)}{\partial t_0}\alpha(t_0)\right\} \times \\ &\quad exp\left\{\alpha^T(t_0)\Psi(y_l, t_0)\right\} f_l(y_l)\, dy_l + \lambda\frac{\partial \rho(t_0)}{\partial t_0} \\ &= 0.\end{aligned}$$

Applying the definition of $h$, the last equation simplifies to

$$nA^T(t_0)\alpha(t_0) + \lambda\frac{\partial \rho(t_0)}{\partial t_0} = 0$$

which is equivalent to our conditions for the proportionality stated in Theorem 3.1. In other words, the two approaches are using the same piece of information. In addition, it shows that our assumption of the existence of a unique $h$ is equivalent to that $g_2(t_0)$ is conditionally unimodal subject to the constraint $\rho(t_0) = \rho_0$.

# Chapter 4

# Some applications

## 4.1 Overview

In this chapter, we demonstrate the performance of $G_p$ through some numerical examples. Specifically, we apply the approximation to the models and estimators which are defined in Section 2.2.

The distributions of $G_p$ for $\hat{\rho} = \rho(\hat{\eta})$ are evaluated by numerical integrations of the density approximation given in (3.11). In the multiple regression example, the three adjustments proposed in Section 3.5 are implemented. Replacing $t_0$ by $t_\mu$ in (3.12), (3.13) and (3.14) for the expectation $E_h[T]$, we denote the adjusted approximations by $G_{p1}$, $G_{p2}$ and $G_{p3}$ respectively.

For comparison, we simulate the true distributions and compute the asymptotic distributions of the function $\hat{\rho}$. In addition, we examine how a linear approximation for $\hat{\rho}$ performs. Specifically, we consider the distributions of the linear function $G$ under $f$. The asymptotic approximation and the linear approximation are denoted by $\hat{\rho}_{asy}$ and $G_f$ respectively. In our examples, the distributions of $\hat{\rho}$ and $G_f$ are based on 100,000 simulations.

Basically, we examine the performances of the approximations for the location-scale and the multiple regression problems under four error distributions, namely, the standard normal ($Z$), the standard $t$ with three degrees of freedom ($t_3$), a contaminated normal ($C_N \equiv .9Z + .1N(0, 100)$) and the standard Cauchy ($t_1$). These represent a wide range of distributions from both theoretical and practical point of view.

To generate the numerical results, we rely on two computer libraries, namely, NAG and ROBETH. Specifically, for the simulations, we call the subroutines G05CAF, G05DDF, and G05DJF in NAG to generate uniform, normal, and $t$ random numbers respectively, and then the subroutines LYHALG and RYNALG in ROBETH to solve for the location-scale and the multiple regression estimates respectively. For our $G_p$, we basically use the subroutine C05NBF in NAG to solve the non-linear system for the required $\alpha$ and $t_0$, and the subroutine D01AJF, also in NAG, for one-dimensional integrations.

In Sections 4.2 and 4.3, we derive specific formulae for the approximations for the location-scale and the multiple regression problems respectively, and comment on the individual results. A general discussion on the approximations is given in Section 4.4. The discussion includes an application of $G_p$ to a Mallows-type estimator. The result can be used as an indicator for further studies. Finally, Section 4.5 summarizes the numerical results which are generated in Sections 4.2, 4.3 and 4.4.

## 4.2 Case 1: Location-scale

This section considers the location-scale problem which is defined in Section 2.2. Recall that the location-scale model is

$$Y_l = \theta + \sigma\varepsilon_l, \quad l = 1, \ldots, n,$$

and the score functions are

$$\Psi_{1l}(Y_l, \eta) = \Psi_c\left(\frac{Y_l - \theta}{\sigma}\right), \quad \Psi_{2l}(Y_l, \eta) = \Psi_c^2\left(\frac{Y_l - \theta}{\sigma}\right) - \beta.$$

Due to the equivariance and the invariance of the estimators, the choice of $\eta_0$ is not important. In the simulation, we take $(\theta_0, \sigma_0) = (0, 1)$. Therefore, the asymptotic result of $\hat{\theta}$ in (2.4) gives

$$\hat{\theta}_{asy} \sim N\left(0, \frac{1}{n}\frac{E_f[\Psi_c^2(\varepsilon_1)]}{\{E_f[I_c(\varepsilon_1)]\}^2}\right),$$

where $I_c(x)$ equals 1 if $|x| < c$, and 0 otherwise. For the reason we will see immediately, we choose $c = 1.345$. With this choice of $c$, the asymptotic variance of $\hat{\theta}$ is

| $f_\varepsilon$ | $Z$ | $t_3$ | $C_N$ | $t_1$ |
|---|---|---|---|---|
| $\dfrac{E_f[\Psi_c^2(\varepsilon_1)]}{\{E_f[I_c(\varepsilon_1)]\}^2}$ | 1.0526 | 1.5565 | 1.4351 | 2.8425 |

and therefore the asymptotic efficiency of $\hat{\theta}$ relative to the arithmetic mean $\overline{Y}$ under $Z$ is .9500. Denoting the standard normal distribution and density functions by $\Phi$ and $\varphi$ respectively, we choose

$$\beta = \int_{-\infty}^{\infty} \Psi_c^2(r)\, d\Phi(r) = 1 - 2\left\{c\varphi(c) + (1 - c^2)\,\Phi(-c)\right\}$$

so that

$$E_\varphi[\Psi_c(Y_1)] = 0 \quad \text{and} \quad E_\varphi[\Psi_c^2(Y_1) - \beta] = 0,$$

and therefore the estimator $\hat{\eta} = (\hat{\theta}, \hat{\sigma})$ is Fisher consistent (see Hampel et al., 1986, page 102). In addition, when $c = \infty$, we have $\beta = 1$, and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{l=1}^{n}(Y_l - \overline{Y})^2$$

is the maximum likelihood estimator of $\sigma^2$ under $Z$. For our examples, we have $\beta = .7102$ with $c = 1.345$.

We have $p = 2$ parameters to be estimated. To demonstrate the performance of the approximations for small samples, we take sample sizes $n$ of 10 and 20, which give the ratio $p : n$ of $1 : 5$ and $1 : 10$ respectively.

With the above setting, we evaluate the quantiles for the estimators $\hat{\theta}$, $\hat{\sigma}$ and their approximations $\hat{\rho}_{asy}$, $G_f$ and $G_p$. Recall that the basis of $G_f$ and $G_p$ is the linear function $G = \rho(t_0) + \overline{G}$ for some chosen $t_0$. Details for $\overline{G}$ can be found in Section 2.5. For our present setting, we have

$$G_f = \frac{1}{n}\sum_{l=1}^{n}\frac{t_\sigma}{E_f[I_c(r_l)]}\Psi_c(r_l) \quad \text{for} \quad \hat{\rho} = \hat{\theta}$$

and

$$G_f = t_\sigma + \frac{1}{n}\sum_{l=1}^{n}\frac{t_\sigma}{2E_f[r_l^2 I_c(r_l)]}\left\{\Psi_c^2(r_l) - \beta\right\} \quad \text{for} \quad \hat{\rho} = \hat{\sigma},$$

where $r_l = t_\sigma^{-1}Y_l$ and $t_0 = (t_\theta, t_\sigma) = (0, t_\sigma)$ satisfies the condition (3.8) under $f$. Note that $t_\sigma$ does not equal to $\sigma_0$ in general. Numerical results are reported in Tables 4.1a and 4.1b for $\hat{\theta}$, Tables 4.2a. 4.2b and Figures 4.1a, 4.1b for $\hat{\sigma}$ in Section 4.5. We use different vertical scales for Figures 4.1a and 4.1b to give better comparisons between $G_f$ and $G_p$ under different error distributions.

Consider the performance for $\hat{\theta}$. The three approximations show basically no difference when the data are normally distributed. Table 4.1a shows that all three approximations give excellent results under $Z$ even when the sample size is just 10. When $n$ is increased to 20, the approximations, except for very small discrepancies found in $G_f$, match the true distribution perfectly.

The situation is slightly changed under $t_3$. Generally, $\hat{\theta}_{asy}$ and $G_f$ behave more or less the same. When $n = 10$, they both give good approximations up to about the ninety-ninth percentile but show larger inaccuracies farther out in the tails. When the sample size is doubled, the two approximations are improving, but some signs of inadequacy can still be found. On the other hand, $G_p$ provides very accurate approximations when $n = 10$ and is almost perfect when $n = 20$. Nevertheless, the overall differences among the three approximations are not yet very significant.

When we go to Table 4.1b, the difference becomes obvious. The numerical results under $C_N$ show that $\hat{\theta}_{asy}$ and $G_f$ are very much alike. They work reasonably well up to the ninety-fifth percentile and the discrepancy grows dramatically thereafter. Increasing sample size does help and makes the two approximations more acceptable. With such a strong contaminated distribution, $G_p$ still works very well even in the tails. There are slight discrepancies found in the far ends of $G_p$, but the approximation is still satisfactory.

Finally, from the results under $t_1$, the performances of all three approximations become distinguishable. The inadequacy of $\hat{\theta}_{asy}$ shows up in the interquartile range and never catches up to the true distribution. $G_f$ on the other hand is very accurate at least within the interquartile range when $n = 10$ and up to the ninetieth percentile when $n = 20$. On the other hand, the performance of $G_p$ is consistently good and

very similar to that under $C_N$.

We now consider the performances of the approximations for $\hat{\sigma}$. We have demonstrated the performance of the asymptotic approximation for a symmetric estimator, $\hat{\theta}$. For an asymmetric estimator such as $\hat{\sigma}$, we do not expect that the asymptotic normality will be able to compete well. In fact, the numerical results show that the true distributions of $\hat{\sigma}$ are generally highly skewed to the right and are nowhere close to a normal distribution. In addition, $G_f$ has shown its superiority over $\hat{\theta}_{asy}$. For these reasons, we consider only the performance of $G_f$ and $G_p$.

From the results in Tables 4.2a and 4.2b, we can see that the performance of $G_f$ and $G_p$ for $\hat{\sigma}$ is very similar to that for $\hat{\theta}$. Rather than give detailed comments for each density, we summarize the overall results and make some general comments on the approximations.

We observe that the distributions of $\hat{\sigma}$ are increasingly asymmetric in the order of the underlying distributions $Z$, $t_3$, $C_N$ and $t_1$. Both $G_f$ and $G_p$ give very good results under $Z$. In addition, $G_p$ provides consistent approximations under all four error distributions.

Generally, $G_f$ works well around the center but tends to generate more symmetric and shorter distributions than $\hat{\sigma}$. In the extreme case when the error is under $t_1$, $G_f$ has approximately one percent of distribution located on the negative region. This is possibly caused by using only the linear term in a Taylor series expansion for approximation. $G_p$ generally improves a lot from $G_f$ and from $n = 10$ to $n = 20$. Nevertheless, we observe that there are some discrepancies between the approximations and the true distribution. We use some diagrams to illustrate the situation.

Figures 4.1a and 4.1b consist of some QQ-plots for $Q_{G_f}$, $Q_{G_p}$ and $Q_{\hat{\sigma}}$, the quantiles

of $G_f$, $G_p$ and $\hat{\sigma}$ respectively. Specifically, it gives $Q_{G_f} - Q_{\hat{\sigma}}$ versus $Q_{\hat{\sigma}}$ and $Q_{G_p} - Q_{\hat{\sigma}}$ versus $Q_{\hat{\sigma}}$ when $n = 10$. In brief, $G_f$ captures the shape around the center and $G_p$ gives in addition consistent approximations up to at least the median. The consistent discrepancy between the distributions of $\hat{\sigma}$ and $G_p$ suggests that a small adjustment to the centering constraints may be useful. We postpone such an adjustment to the next example.

## 4.3 Case 2: Multiple regression

We have demonstrated the performance of $G_p$ and other approximations in a location-scale problem, we now consider a non-identically distributed situation. Recall from Section 2.2 that the multiple regression model is

$$Y_l = X_l^T\theta + \sigma\varepsilon_l, \quad l = 1, \ldots, n,$$

where $X$ is an $n$ by $p-1$ fixed design, and the score functions are

$$\Psi_{jl}(Y_l, \eta) = \Psi_c\left(\frac{Y_l - X_l^T\theta}{\sigma}\right) X_{lj}, \quad j = 1, \ldots, p-1,$$

$$\Psi_{pl}(Y_l, \eta) = \Psi_c^2\left(\frac{Y_l - X_l^T\theta}{\sigma}\right) - \beta.$$

For this example, we arbitrarily take $\theta_0 = (1, 1, -1, 2)$ and $\sigma_0 = 1$. In addition, we set $n = 20$ and $p = 5$, which gives the ratio of $n$ to $p$ of 4. The design matrix $X = X_{20\times 4}$ is generated from a uniform distribution $U(0, 1)$ except for the first column which equals 1's such that

$$diag\left\{(X^TX)^{-1}\right\} = (.56, .62, .61, .51)$$

and

$$\begin{aligned} diag\left\{X(X^TX)^{-1}X^T\right\} \;=\; & (.19, .20, .31, .17, .15, .12, .17, .16, .22, .32, \\ & .22, .22, .07, .24, .11, .27, .24, .23, .19, .22). \end{aligned}$$

By using the rule of thumb $2(p-1)/n$ (see Hoaglin and Welsch, 1978), which equals .4 in our case, we do not have any obvious potential influence points in the design. Lastly, with similar reasons as in the last example, we choose $c = 1.345$ and

$$\beta = \frac{n-p+1}{n}\int_{-\infty}^{\infty}\Psi_c^2(r)\,d\Phi(r) = .5681.$$

From the asymptotic result of $\hat{\theta}$ in (2.5) and the diagonal of $(X^TX)^{-1}$ above, we see that the asymptotic behaviours of $\hat{\theta}_k$, $k = 1,\ldots,4$, are very similar. Numerical comparisons are based on the estimators $\hat{\theta}_3$ and $\hat{\sigma}$. In particular, we have

$$\hat{\theta}_{3\,asy} \sim N\left(\theta_{30}, \frac{E_f[\Psi_c^2(\varepsilon_1)]}{\{E_f[I_c(\varepsilon_1)]\}^2}(X^TX)^{-1}_{33}\right),$$

where $I_c(x)$ equals 1 if $|x| < c$, and 0 otherwise, and the variance of $\hat{\theta}_{3\,asy}$ is given by

| $f_\varepsilon$ | $Z$ | $t_3$ | $C_N$ | $t_1$ |
|---|---|---|---|---|
| $\frac{E_f[\Psi_c^2(\varepsilon_1)]}{\{E_f[I_c(\varepsilon_1)]\}^2}(X^TX)^{-1}_{33}$ | .6435 | .9516 | .8774 | 1.7378 |

For the distributions of $G_f$ and $G_p$, we need the linear approximation $G$. The construction of it is similar to that for the location-scale problem. For instance, consider the approximation $G_f$ for $\hat{\theta}_3$. Under our setting, $t_0 = (\theta_0, t_\sigma)$ that satisfies the condition (3.8) under $f$,

$$A = -\frac{1}{nt_\sigma}E_f\begin{bmatrix} X^TDX & X^TDr \\ 2r^TDX & 2r^TDr \end{bmatrix} = -\frac{1}{nt_\sigma}\begin{bmatrix} X^TE_f[D]X & 0 \\ 0 & 2E_f[r^TDr] \end{bmatrix}$$

and

$$B = nt_\sigma\begin{bmatrix} \{X^TE_f[D]X\}^{-1} & 0 \\ 0 & \{2E_f[r^TDr]\}^{-1} \end{bmatrix},$$

where $r = (r_1,\ldots,r_n)$, $D$ is an $n$ by $n$ diagonal matrix with diagonal elements $I_c(r_l)$,

$$r_l = \frac{Y_l - X_l^T\theta_0}{t_\sigma},\quad l = 1,\ldots,n.$$

Therefore we have

$$G_f = \theta_{30} + \sum_{l=1}^{n}\sum_{j=1}^{p-1} t_\sigma \left\{X^T E_f[D]X\right\}^{-1}_{3j} \Psi_c(r_l)X_{lj}.$$

Table 4.3 summarizes the numerical results for $\hat{\theta}_3 - \theta_{30}$. In general, the situation is very similar to, but slightly more contrasting than that in the location-scale problem. All three approximations $\hat{\theta}_{3asy}$, $G_f$ and $G_p$ give excellent results under $Z$. $G_f$ and $\hat{\theta}_{3asy}$ behave very much alike under $t_3$ and $C_N$. Basically, both of them provide very good approximations under $t_3$ except in the tails, but quite unacceptable results under $C_N$ and $t_1$. Nevertheless, $G_f$ seems to be marginally better than $\hat{\theta}_{3asy}$. In addition, $G_f$ presents some definite advantages over $\hat{\theta}_{3asy}$ around the center under $t_1$.

On the other hand, $G_p$ works consistently well under $t_3$. Under the last two distributions where $G_f$ deviates substantially from the true distribution, $G_p$ reduces the discrepancy by approximately two thirds and generally gives us a fair idea of the true situation. We realize that there is room for $G_p$ to be improved. For the next two estimators, we implement the proposed adjustments from Section 3.5 and determine if they are helpful.

Tables 4.4a and 4.4b summarize the quantiles of the true distribution and the approximations $G_f$, $G_p$, $G_{p1}$, $G_{p2}$ and $G_{p3}$ for $\hat{\sigma}$. Basically, $G_f$ and $G_p$ behave more or less the same as their counterparts in the location-scale problem. $G_{p1}$ improves the approximation around the center but distorts in the tails. $G_{p2}$ works extremely well under $Z$ and very well under $t_3$. It improves the approximation a bit on one end and distorts it a bit on the other end under $C_N$ and $t_1$. Compared to $G_{p2}$, $G_{p3}$ is more consistent. It always improves the low end and distorts the other end. Based on these results, we do not claim to have found a reasonable adjustment in general. Nevertheless, the benefit of using these adjustments is clear in the following practical

situation.

Recall that we have shown in Section 2.2 that the function

$$\rho(\hat{\eta}) = \rho\left(\frac{\hat{\theta}_3 - \theta_{30}}{\hat{\sigma}}\right)$$

is location-scale invariant. Therefore, a statistic of this form can be used for inferential purposes. For instance, a commonly used non-linear function of this form for making statistical inferences on $\theta_3$ is a studentized version of $\hat{\theta}_3$, that is,

$$\rho(\hat{\eta}) = \frac{\hat{\theta}_3 - \theta_{30}}{\gamma\hat{\sigma}},$$

where $\gamma^2$ is the variance of $\hat{\theta}_{3asy}$. Now $\hat{\rho}_{asy}$ has the standard normal distribution. In addition, we may expect from the classical theory that a $t$ distribution will give better approximations than $\hat{\rho}_{asy}$. We evaluate the $t$ distribution with sixteen degrees of freedom and denote it by $\hat{\rho}_{t_{16}}$ for comparison. For our $G_p$, the derivation of the linear approximation $G$ for the non-linear ratio is similar to the previous constructions. We leave the details to the next section where a more general situation will be considered.

Tables 4.5a and 4.5b report the tail areas for the function $\hat{\rho}$ and the approximations. In brief, $\hat{\rho}_{asy}$ gives fair approximations under $Z$ and conservative results under other distributions. In contrast with what we think, $\hat{\rho}_{t_{16}}$ gives an improvement over $\hat{\rho}_{asy}$ only for the case under $Z$. In general, it gives even more conservative results than $\hat{\rho}_{asy}$. Field (1982) observed that the $t$ distribution with a reduced degree of freedom can give a better agreement with the true distribution.

For the performance of $G_p$, it generally gives rough but consistent approximations. $G_{p1}$ improves the approximation around the center. $G_{p2}$ works very well under $Z$, improves the approximation around the center under $t_3$ and $C_N$, but is getting worse under $t_1$. $G_{p3}$ consistently improves the approximations under all four error distributions.

Figures 4.2a and 4.2b plot $L_{approx} - L_{exact}$ versus $L_{exact}$ for the distributions from Tables 4.5a and 4.5b in logistic scale. This provides a better picture of their performances in the tails. In brief, the plots show the general inadequacy of the asymptotic results and the overall accuracy of $G_p$ and its variations.

Comparing the performance of $G_p$ for $\hat{\sigma}$ and that for the studentized $t$-ratio, it seems quite clear that the unknown scale is a major problem and supports our effort for improving the approximations for $\hat{\sigma}$. Moreover, it is probably more important to get $\hat{\sigma}$ correct in the lower tail.

To conclude, if we use the studentized ratio for testing a hypothesis under $Z$, $G_{p2}$ is the simplest one and gives the best approximation among the three variations. However, if our objective is to study the behaviour of the ratio, $G_{p3}$ would be a sensible choice since it gives more stable approximations.

## 4.4 Discussion

From the numerical results of the two examples, we find that $\hat{\theta}_{asy}$ is an excellent approximation under $Z$ and reasonably accurate under $t_3$. $G_f$ is at least as good as $\hat{\theta}_{asy}$ and in some situations demonstrates its superiority over the latter one. In the worst case, that is under $t_1$, $G_f$ still provides a very accurate approximation around the expected value, which is exactly what our $G_p$ is built upon. $G_p$ provides excellent approximations for $\hat{\theta}$ under all four error distributions and no adjustment is needed at all. For the studentized $t$-ratio, the $t$ distribution works very well under $Z$ but it becomes inadequate under the other distributions. $G_p$ is consistent but the numerical approximations deteriorate as we move out into the tail. We have proposed some adjustments and have demonstrated their usefulness.

To summarize, when the distribution of a simple estimator such as $\hat{\theta}$ under $Z$ is needed, the asymptotic result is undoubtedly the best choice. It is simple and accurate. Otherwise, if only the central distribution is what we need, such as constructing a confidence interval with a moderate level, say 80% or 90%, $G_f$ can be used to give reliable results over various underlying distributions. However, if the tail distribution is our main concern, or if we need to study and compare the behaviours of different estimators, $G_p$ is clearly the best alternative among the three.

To conclude this chapter, we give a simple example to demonstrate the consistent performance of $G_p$ in a more general problem. The problem is similar to our second example. We replace the Huber-type score functions by the Mallows-type ones using the optimal standardized weight $W_l$ (see Hampel et al., 1986, page 321), $l = 1, \ldots, n$. The score functions are

$$\Psi_{jl}(Y_l, \eta) = \Psi_c\left(\frac{Y_l - X_l^T\theta}{\sigma}\right) X_{lj} W_l, \quad j = 1, \ldots, p-1,$$

$$\Psi_{pl}(Y_l, \eta) = \Psi_c^2\left(\frac{Y_l - X_l^T\theta}{\sigma}\right) W_l - \beta.$$

Consider $G_f$ for the studentized $t$-ratio. For the construction of $G$, we have basically the same matrices $A$ and $B$ as for the Huber-type estimators, except that the diagonal elements of $D$ are now replaced by $I_c(r_l)W_l$, $l = 1, \ldots, n$. Since

$$\left.\frac{\partial\rho}{\partial\theta_3}\right|_{\eta=t_0} = \frac{1}{\gamma t_\sigma}, \quad \left.\frac{\partial\rho}{\partial\sigma}\right|_{\eta=t_0} = -\frac{\theta_{30} - \theta_{30}}{\gamma t_\sigma^2} = 0,$$

we obtain

$$G_f = \frac{1}{\gamma}\sum_{l=1}^{n}\sum_{j=1}^{p-1}\left\{X^T E_f[D]X\right\}_{3j}^{-1} \Psi_c(r_l) X_{lj} W_l \,.$$

We compute the distributions of the ratio under $Z$ for two designs. The first design has

$$\begin{aligned} diag\left\{X(X^TX)^{-1}X^T\right\} \;=\; & (.25, .28, .24, .23, .19, .17, .25, .18, .12, .15, \\ & .18, .19, .21, .09, .36, .10, .16, .23, .24, .19). \end{aligned}$$

The second design replaces the first two points in the first design to produce influence points and has

$$\begin{aligned} diag\left\{X(X^TX)^{-1}X^T\right\} \;=\; & (.88, .12, .24, .24, .06, .11, .34, .06, .07, .15, \\ & .11, .07, .17, .09, .46, .10, .17, .27, .08, .20). \end{aligned}$$

The weights corresponding to the first and the second designs are

$$W = (.84, .85, .97, .95, 1, 1, .89, 1, 1, 1, 1, 1, 1, 1, .71, 1, 1, .97, 1, 1)$$

and

$$W = (.10, .31, .85, .75, 1, 1, .64, 1, 1, 1, 1, 1, .88, 1, .55, 1, 1, .75, 1, .97)$$

respectively.

Numerical results are summarized in Table 4.6. Basically, the performances of the approximations are very similar to those for the Huber's case. In other words, the asymptotic results are still inadequate, and $G_p$ and its variations still perform consistently well. Adding influence points does not seem to have too much effect on their performances. This encourages us to study the approximations further.

## 4.5 Numerical results

Table 4.1a
Location-scale: Quantiles for $\hat{\theta}$

| | $n = 10$ | | | | $n = 20$ | | | |
|---|---|---|---|---|---|---|---|---|
| $F$ | $\hat{\theta}$ | $\hat{\theta}_{asy}$ | $G_f$ | $G_p$ | $\hat{\theta}$ | $\hat{\theta}_{asy}$ | $G_f$ | $G_p$ |
| | | | | $Z$ | | | | |
| .75 | .22 | .22 | .22 | .22 | .15 | .15 | .15 | .15 |
| .9 | .41 | .42 | .42 | .41 | .29 | .29 | .29 | .29 |
| .95 | .53 | .53 | .53 | .53 | .38 | .38 | .38 | .38 |
| .975 | .63 | .64 | .63 | .63 | .45 | .45 | .45 | .45 |
| .99 | .75 | .75 | .75 | .75 | .53 | .53 | .53 | .53 |
| .995 | .83 | .84 | .83 | .83 | .59 | .59 | .58 | .59 |
| .9975 | .90 | .91 | .89 | .91 | .64 | .64 | .64 | .64 |
| .999 | .99 | 1.00 | .97 | 1.00 | .71 | .71 | .70 | .71 |
| | | | | $t_3$ | | | | |
| .75 | .27 | .27 | .27 | .27 | .19 | .19 | .19 | .19 |
| .9 | .53 | .51 | .52 | .52 | .37 | .36 | .36 | .37 |
| .95 | .68 | .65 | .66 | .67 | .48 | .46 | .47 | .47 |
| .975 | .83 | .77 | .79 | .82 | .57 | .55 | .56 | .57 |
| .99 | 1.01 | .92 | .93 | .99 | .69 | .65 | .66 | .68 |
| .995 | 1.15 | 1.02 | 1.02 | 1.12 | .77 | .72 | .73 | .76 |
| .9975 | 1.28 | 1.11 | 1.10 | 1.25 | .85 | .78 | .79 | .84 |
| .999 | 1.44 | 1.22 | 1.20 | 1.42 | .94 | .86 | .88 | .94 |

Table 4.1b
Location-scale: Quantiles for $\hat{\theta}$

| | $n = 10$ | | | | $n = 20$ | | | |
|---|---|---|---|---|---|---|---|---|
| $F$ | $\hat{\theta}$ | $\hat{\theta}_{asy}$ | $G_f$ | $G_p$ | $\hat{\theta}$ | $\hat{\theta}_{asy}$ | $G_f$ | $G_p$ |
| | | | | $C_N$ | | | | |
| .75 | .26 | .26 | .26 | .25 | .18 | .18 | .18 | .18 |
| .9 | .51 | .49 | .49 | .49 | .35 | .34 | .35 | .34 |
| .95 | .67 | .62 | .63 | .64 | .45 | .44 | .45 | .45 |
| .975 | .83 | .74 | .74 | .78 | .55 | .53 | .53 | .54 |
| .99 | 1.06 | .88 | .88 | 1.03 | .67 | .62 | .62 | .65 |
| .995 | 1.35 | .98 | .97 | 1.36 | .76 | .69 | .69 | .73 |
| .9975 | 1.91 | 1.06 | 1.05 | 1.82 | .85 | .75 | .75 | .81 |
| .999 | 2.77 | 1.17 | 1.14 | 2.54 | .99 | .83 | .82 | .94 |
| | | | | $t_1$ | | | | |
| .75 | .42 | .36 | .43 | .42 | .30 | .25 | .30 | .30 |
| .9 | .90 | .68 | .81 | .90 | .60 | .48 | .57 | .60 |
| .95 | 1.28 | .88 | 1.03 | 1.28 | .81 | .62 | .73 | .81 |
| .975 | 1.71 | 1.04 | 1.22 | 1.70 | 1.03 | .74 | .87 | 1.02 |
| .99 | 2.41 | 1.24 | 1.44 | 2.38 | 1.32 | .88 | 1.03 | 1.30 |
| .995 | 3.08 | 1.37 | 1.59 | 3.00 | 1.55 | .97 | 1.14 | 1.54 |
| .9975 | 3.95 | 1.50 | 1.72 | 3.75 | 1.80 | 1.06 | 1.23 | 1.78 |
| .999 | 5.28 | 1.65 | 1.88 | 4.99 | 2.15 | 1.17 | 1.36 | 2.14 |

Table 4.2a
Location-scale: Quantiles for $\hat{\sigma}$

| | $n$=10 | | | $n$=20 | |
|---|---|---|---|---|---|
| $F$ | $\hat{\sigma}$ | $G_f$ | $G_p$ | $\hat{\sigma}$ | $G_p$ |
| | | | $Z$ | | |
| .001 | .25 | .27 | .30 | .43 | .46 |
| .005 | .34 | .36 | .39 | .51 | .54 |
| .01 | .38 | .41 | .44 | .55 | .58 |
| .05 | .51 | .56 | .58 | .66 | .69 |
| .1 | .60 | .65 | .66 | .72 | .76 |
| .5 | .92 | .99 | .99 | .96 | .99 |
| .9 | 1.29 | 1.36 | 1.35 | 1.22 | 1.25 |
| .95 | 1.40 | 1.47 | 1.46 | 1.29 | 1.32 |
| .99 | 1.61 | 1.67 | 1.67 | 1.44 | 1.47 |
| .995 | 1.69 | 1.74 | 1.75 | 1.49 | 1.52 |
| .999 | 1.87 | 1.89 | 1.92 | 1.61 | 1.64 |
| | | | $t_3$ | | |
| .001 | .28 | .17 | .33 | .49 | .51 |
| .005 | .37 | .28 | .43 | .58 | .60 |
| .01 | .43 | .36 | .48 | .62 | .65 |
| .05 | .59 | .57 | .64 | 76 | .78 |
| .1 | .69 | .69 | .74 | .83 | .86 |
| .5 | 1.11 | 1.18 | 1.17 | 1.15 | 1.18 |
| .9 | 1.71 | 1.71 | 1.76 | 1.55 | 1.57 |
| .95 | 1.91 | 1.86 | 1.96 | 1.68 | 1.71 |
| .99 | 2.37 | 2.14 | 2.42 | 1.96 | 1.98 |
| .995 | 2.56 | 2.24 | 2.61 | 2.08 | 2.09 |
| .999 | 3.03 | 2.45 | 3.06 | 2.34 | 2.33 |

Table 4.2b
Location-scale: Quantiles for $\hat{\sigma}$

| | $n$=10 | | | $n$=20 | |
|---|---|---|---|---|---|
| $F$ | $\hat{\sigma}$ | $G_f$ | $G_p$ | $\hat{\sigma}$ | $G_p$ |
| | | | $C_N$ | | |
| .001 | .28 | .21 | .33 | .49 | .51 |
| .005 | .37 | .32 | .43 | .57 | .60 |
| .01 | .42 | .39 | .48 | .61 | .64 |
| .05 | .57 | .58 | .64 | .74 | .77 |
| .1 | .67 | .69 | .74 | .81 | .84 |
| .5 | 1.07 | 1.14 | 1.13 | 1.11 | 1.14 |
| .9 | 1.66 | 1.61 | 1.68 | 1.49 | 1.50 |
| .95 | 1.93 | 1.75 | 1.90 | 1.62 | 1.62 |
| .99 | 3.07 | 2.01 | 2.61 | 1.95 | 1.91 |
| .995 | 4.17 | 2.11 | 3.15 | 2.12 | 2.05 |
| .999 | 6.47 | 2.30 | 5.30 | 2.68 | 2.43 |
| | | | $t_1$ | | |
| .001 | .34 | -.42 | .39 | .60 | .62 |
| .005 | .46 | -.16 | .51 | .72 | .74 |
| .01 | .53 | -.01 | .58 | .78 | .81 |
| .05 | .76 | .42 | .80 | .99 | 1.01 |
| .1 | .90 | .70 | .95 | 1.12 | 1.13 |
| .5 | 1.69 | 1.71 | 1.72 | 1.73 | 1.73 |
| .9 | 3.36 | 2.83 | 3.28 | 2.75 | 2.71 |
| .95 | 4.20 | 3.16 | 4.02 | 3.17 | 3.10 |
| .99 | 6.91 | 3.77 | 6.14 | 4.22 | 4.07 |
| .995 | 8.54 | 3.97 | 7.28 | 4.72 | 4.52 |
| .999 | 13.60 | 4.41 | 10.64 | 6.08 | 5.68 |

Table 4.3
Regression (Huber's): Quantiles for $\hat{\theta}_3$

| $F$ | $\hat{\theta}_3$ | $\hat{\theta}_{3\,asy}$ | $G_f$ | $G_p$ | $\hat{\theta}_3$ | $\hat{\theta}_{3\,asy}$ | $G_f$ | $G_p$ |
|---|---|---|---|---|---|---|---|---|
| | $Z$ | | | | $t_3$ | | | |
| .75 | .53 | .54 | .54 | .53 | .70 | .66 | .68 | .68 |
| .9 | 1.01 | 1.03 | 1.01 | 1.01 | 1.34 | 1.25 | 1.30 | 1.31 |
| .95 | 1.31 | 1.32 | 1.30 | 1.30 | 1.75 | 1.60 | 1.66 | 1.70 |
| .975 | 1.56 | 1.57 | 1.56 | 1.55 | 2.14 | 1.91 | 1.99 | 2.05 |
| .99 | 1.86 | 1.87 | 1.83 | 1.84 | 2.58 | 2.27 | 2.36 | 2.48 |
| .995 | 2.05 | 2.07 | 2.02 | 2.04 | 2.91 | 2.51 | 2.60 | 2.79 |
| .9975 | 2.24 | 2.25 | 2.20 | 2.22 | 3.23 | 2.74 | 2.81 | 3.09 |
| .999 | 2.46 | 2.48 | 2.41 | 2.45 | 3.63 | 3.01 | 3.09 | 3.47 |
| | $C_N$ | | | | $t_1$ | | | |
| .75 | .70 | .63 | .65 | .64 | 1.28 | .89 | 1.17 | 1.17 |
| .9 | 1.37 | 1.20 | 1.24 | 1.24 | 2.66 | 1.69 | 2.22 | 2.39 |
| .95 | 1.83 | 1.54 | 1.59 | 1.63 | 3.70 | 2.17 | 2.85 | 3.28 |
| .975 | 2.31 | 1.84 | 1.89 | 1.99 | 4.83 | 2.58 | 3.39 | 4.21 |
| .99 | 3.09 | 2.18 | 2.23 | 2.51 | 6.52 | 3.07 | 4.03 | 5.55 |
| .995 | 3.83 | 2.41 | 2.47 | 2.98 | 8.04 | 3.40 | 4.44 | 6.69 |
| .9975 | 4.73 | 2.63 | 2.69 | 3.59 | 9.88 | 3.70 | 4.84 | 7.96 |
| .999 | 6.10 | 2.89 | 2.95 | 4.66 | 12.45 | 4.07 | 5.29 | 9.89 |

Table 4.4a
Regression (Huber's): Quantiles for $\hat{\sigma}$

| $F$ | $\hat{\sigma}$ | $G_f$ | $G_p$ | $G_{p1}$ | $G_{p2}$ | $G_{p3}$ |
|---|---|---|---|---|---|---|
| | | | $Z$ | | | |
| .001 | .45 | .62 | .60 | .59 | .45 | .52 |
| .005 | .54 | .70 | .68 | .66 | .54 | .60 |
| .01 | .58 | .74 | .73 | .69 | .58 | .63 |
| .05 | .71 | .86 | .85 | .79 | .71 | .75 |
| .1 | .78 | .93 | .92 | .84 | .78 | .81 |
| .5 | 1.04 | 1.19 | 1.19 | 1.04 | 1.04 | 1.05 |
| .9 | 1.33 | 1.49 | 1.48 | 1.26 | 1.33 | 1.30 |
| .95 | 1.41 | 1.57 | 1.56 | 1.33 | 1.41 | 1.38 |
| .99 | 1.58 | 1.74 | 1.72 | 1.45 | 1.57 | 1.52 |
| .995 | 1.64 | 1.79 | 1.78 | 1.49 | 1.63 | 1.57 |
| .999 | 1.76 | 1.93 | 1.90 | 1.59 | 1.76 | 1.68 |
| | | | $t_3$ | | | |
| .001 | .54 | .59 | .67 | .67 | .55 | .58 |
| .005 | .64 | .69 | .77 | .76 | .66 | .67 |
| .01 | .69 | .76 | .82 | .81 | .71 | .72 |
| .05 | .84 | .94 | .99 | .95 | .87 | .86 |
| .1 | .94 | 1.05 | 1.08 | 1.03 | .97 | .94 |
| .5 | 1.32 | 1.47 | 1.46 | 1.35 | 1.35 | 1.27 |
| .9 | 1.81 | 1.92 | 1.95 | 1.76 | 1.84 | 1.70 |
| .95 | 1.98 | 2.06 | 2.11 | 1.89 | 2.00 | 1.84 |
| .99 | 2.34 | 2.32 | 2.46 | 2.16 | 2.35 | 2.13 |
| .995 | 2.50 | 2.41 | 2.60 | 2.27 | 2.49 | 2.25 |
| .999 | 2.85 | 2.63 | 2.92 | 2.51 | 2.81 | 2.53 |

Table 4.4b
Regression (Huber's): Quantiles for $\hat{\sigma}$

| $F$ | $\hat{\sigma}$ | $G_f$ | $G_r$ | $G_{p1}$ | $G_{p2}$ | $G_{p3}$ |
|---|---|---|---|---|---|---|
| | | | $C_N$ | | | |
| .001 | .52 | .58 | .66 | .68 | .57 | .57 |
| .005 | .62 | .70 | .76 | .76 | .67 | .66 |
| .01 | .67 | .75 | .81 | .81 | .72 | .71 |
| .05 | .82 | .92 | .96 | .94 | .87 | .84 |
| .1 | .91 | 1.02 | 1.05 | 1.02 | .96 | .91 |
| .5 | 1.28 | 1.40 | 1.40 | 1.39 | 1.30 | 1.22 |
| .9 | 1.90 | 1.82 | 1.88 | 2.19 | 1.78 | 1.62 |
| .95 | 2.25 | 1.95 | 2.07 | 2.66 | 1.98 | 1.77 |
| .99 | 3.50 | 2.19 | 2.62 | 3.90 | 2.53 | 2.20 |
| .995 | 4.12 | 2.27 | 2.95 | 4.36 | 2.86 | 2.45 |
| .999 | 5.61 | 2.47 | 4.11 | 5.03 | 4.01 | 3.31 |
| | | | $t_1$ | | | |
| .001 | .70 | .13 | .83 | .90 | 1.00 | .71 |
| .005 | .87 | .40 | .98 | 1.06 | 1.15 | .85 |
| .01 | .95 | .55 | 1.07 | 1.14 | 1.23 | .92 |
| .05 | 1.23 | 1.02 | 1.33 | 1.42 | 1.50 | 1.15 |
| .1 | 1.40 | 1.30 | 1.50 | 1.60 | 1.67 | 1.29 |
| .5 | 2.31 | 2.34 | 2.33 | 2.51 | 2.50 | 1.99 |
| .9 | 4.07 | 3.48 | 3.80 | 4.11 | 3.97 | 3.22 |
| .95 | 4.90 | 3.82 | 4.43 | 4.76 | 4.59 | 3.73 |
| .99 | 7.18 | 4.45 | 6.03 | 6.31 | 6.19 | 5.03 |
| .995 | 8.37 | 4.68 | 6.80 | 6.99 | 6.96 | 5.65 |
| .999 | 12.00 | 5.20 | 8.86 | 8.52 | 9.03 | 7.30 |

Table 4.5a
Regression (Huber's): Tail probabilities for $\hat{\rho} = \frac{\hat{\theta}_3 - \theta_{30}}{\gamma\hat{\sigma}}$

| $1-F$ | std. err. | $\hat{\rho}_{asy}$ | $\hat{\rho}_{t_{16}}$ | $G_p$ | $G_{p1}$ | $G_{p2}$ | $G_{p3}$ |
|---|---|---|---|---|---|---|---|
| | | | | $Z$ | | | |
| .25 | .0014 | .2593 | .2639 | .2225 | .2431 | .2518 | .2501 |
| .1 | .0009 | .1055 | .1145 | .0745 | .0916 | .1032 | .1003 |
| .05 | .0007 | .0489 | .0586 | .0312 | .0411 | .0513 | .0483 |
| .025 | .0005 | .0216 | .0302 | .0133 | .0183 | .0260 | .0235 |
| .01 | .0003 | .0066 | .0124 | .0043 | .0061 | .0108 | .0090 |
| .005 | .0002 | .0024 | .0061 | .0018 | .0025 | .0055 | .0043 |
| .0025 | .0002 | .0007 | .0029 | .0007 | .0010 | .0027 | .0019 |
| .001 | .0001 | .0001 | .0010 | .0002 | .0003 | .0011 | .0007 |
| | | | | $t_3$ | | | |
| .25 | .0014 | .2927 | .2965 | .2239 | .2428 | .2419 | .2505 |
| .1 | .0009 | .1467 | .1545 | .0743 | .0904 | .0915 | .0996 |
| .05 | .0007 | .0851 | .0945 | .0313 | .0410 | .0431 | .0486 |
| .025 | .0005 | .0473 | .0570 | .0127 | .0176 | .0199 | .0229 |
| .01 | .0003 | .0211 | .0295 | .0040 | .0056 | .0073 | .0086 |
| .005 | .0002 | .0103 | .0171 | .0016 | .0021 | .0033 | .0038 |
| .0025 | .0002 | .0048 | .0099 | .0006 | .0008 | .0015 | .0017 |
| .001 | .0001 | .0015 | .0045 | .0002 | .0002 | .0005 | .0006 |

Table 4.5b
Regression (Huber's): Tail probabilities for $\hat{\rho} = \frac{\hat{\theta}_3 - \theta_{30}}{\gamma\hat{\sigma}}$

| $1-F$ | std. err. | $\hat{\rho}_{asy}$ | $\hat{\rho}_{t_{16}}$ | $G_p$ | $G_{p1}$ | $G_{p2}$ | $G_{p3}$ |
|---|---|---|---|---|---|---|---|
| | | | | $C_N$ | | | |
| .25 | .0014 | .2798 | .2839 | .2168 | .2380 | .2327 | .2433 |
| .1 | .0009 | .1320 | .1402 | .0698 | .0860 | .0842 | .0940 |
| .05 | .0007 | .0740 | .0836 | .0296 | .0386 | .0394 | .0460 |
| .025 | .0005 | .0411 | .0508 | .0129 | .0172 | .0189 | .0229 |
| .01 | .0003 | .0176 | .0257 | .0042 | .0055 | .0071 | .0088 |
| .005 | .0002 | .0085 | .0148 | .0017 | .0022 | .0033 | .0041 |
| .0025 | .0002 | .0037 | .0083 | .0007 | .0008 | .0015 | .0019 |
| .001 | .0001 | .0013 | .0042 | .0002 | .0002 | .0006 | .0007 |
| | | | | $t_1$ | | | |
| .25 | .0014 | .3350 | .3378 | .2233 | .2408 | .2071 | .2497 |
| .1 | .0009 | .2087 | .2147 | .0721 | .0865 | .0585 | .0974 |
| .05 | .0007 | .1497 | .1574 | .0302 | .0386 | .0219 | .0476 |
| .025 | .0005 | .1086 | .1175 | .0127 | .0168 | .0082 | .0232 |
| .01 | .0003 | .0705 | .0802 | .0039 | .0052 | .0022 | .0087 |
| .005 | .0002 | .0489 | .0586 | .0015 | .0019 | .0007 | .0039 |
| .0025 | .0002 | .0345 | .0439 | .0006 | .0007 | .0003 | .0019 |
| .001 | .0001 | .0199 | .0283 | .0002 | .0001 | .0001 | .0006 |

Table 4.6
Regression (Mallow's): Tail probabilities for $\hat{\rho} = \frac{\hat{\theta}_3 - \theta_{30}}{\gamma\hat{\sigma}}$ under $Z$

| $1-F$ | std. err. | $\hat{\rho}_{asy}$ | $\hat{\rho}_{t_{16}}$ | $G_p$ | $G_{p1}$ | $G_{p2}$ | $G_{p3}$ |
|---|---|---|---|---|---|---|---|
| | | | | Design 1 | | | |
| .25 | .0014 | .2604 | .2649 | .2237 | .2443 | .2525 | .2510 |
| .1 | .0009 | .1045 | .1135 | .0738 | .0908 | .1019 | .0991 |
| .05 | .0007 | .0498 | .0595 | .0318 | .0420 | .0518 | .0489 |
| .025 | .0005 | .0216 | .0302 | .0133 | .0184 | .0259 | .0234 |
| .01 | .0003 | .0061 | .0117 | .0041 | .0057 | .0101 | .0085 |
| .005 | .0002 | .0021 | .0057 | .0017 | .0023 | .0051 | .0040 |
| .0025 | .0002 | .0006 | .0027 | .0007 | .0009 | .0025 | .0018 |
| .001 | .0001 | .0002 | .0012 | .0003 | .0003 | .0012 | .0008 |
| | | | | Design 2 | | | |
| .25 | .0014 | .2639 | .2683 | .2271 | .2434 | .2514 | .2508 |
| .1 | .0009 | .1070 | .1160 | .0760 | .0892 | .0998 | .0981 |
| .05 | .0007 | .0513 | .0610 | .0331 | .0408 | .0500 | .0482 |
| .025 | .0005 | .0226 | .0312 | .0142 | .0179 | .0248 | .0232 |
| .01 | .0003 | .0070 | .0129 | .0048 | .0060 | .0102 | .0090 |
| .005 | .0002 | .0025 | .0064 | .0021 | .0025 | .0051 | .0043 |
| .0025 | .0002 | .0009 | .0034 | .0010 | .0011 | .0028 | .0022 |
| .001 | .0001 | .0002 | .0012 | .0003 | .0003 | .0011 | .0008 |

Figure 4.1a
Location-scale: QQ-plots for $\hat{\sigma}$

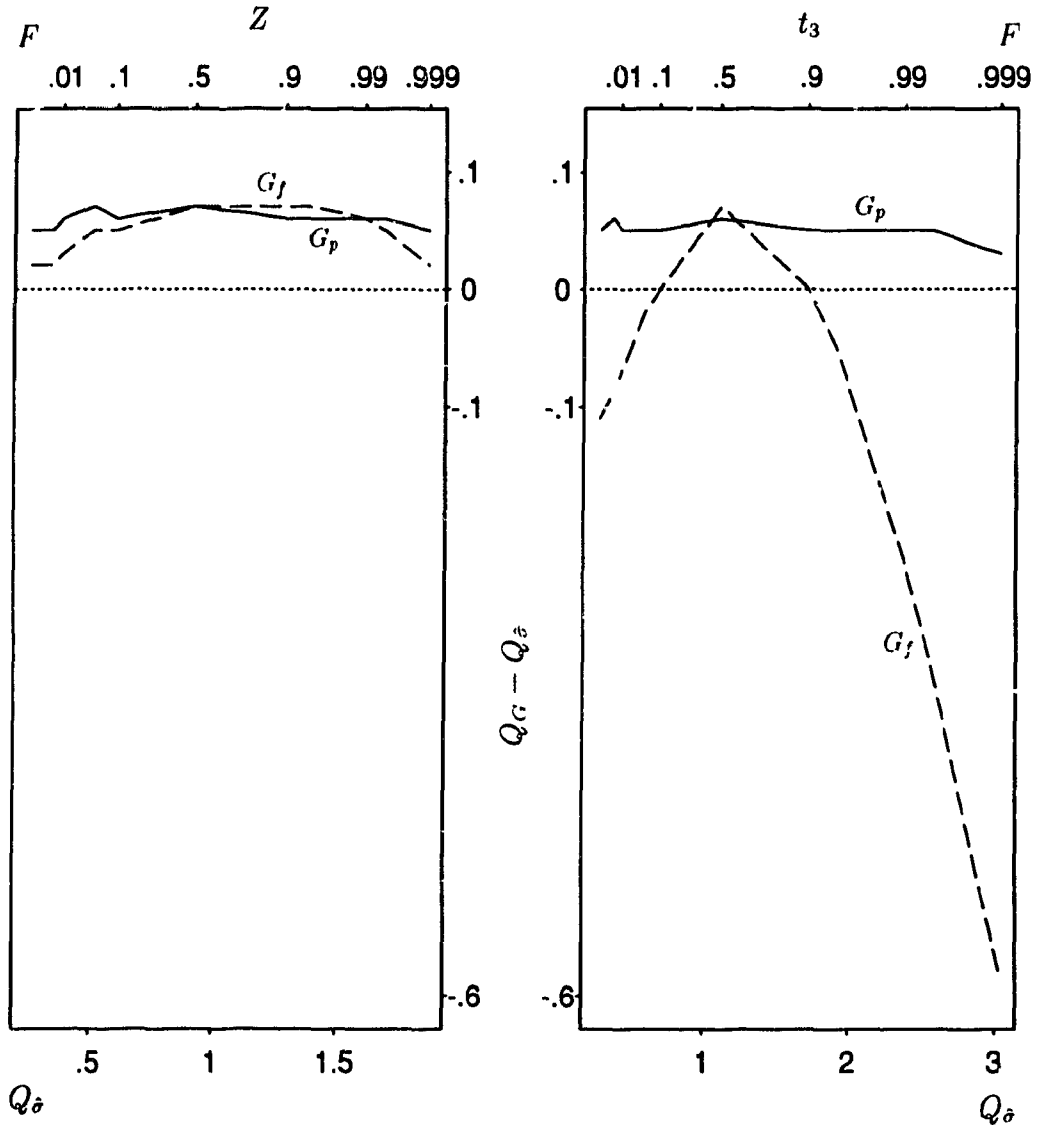


Figure 4.1b
Location-scale: QQ-plots for $\hat{\sigma}$

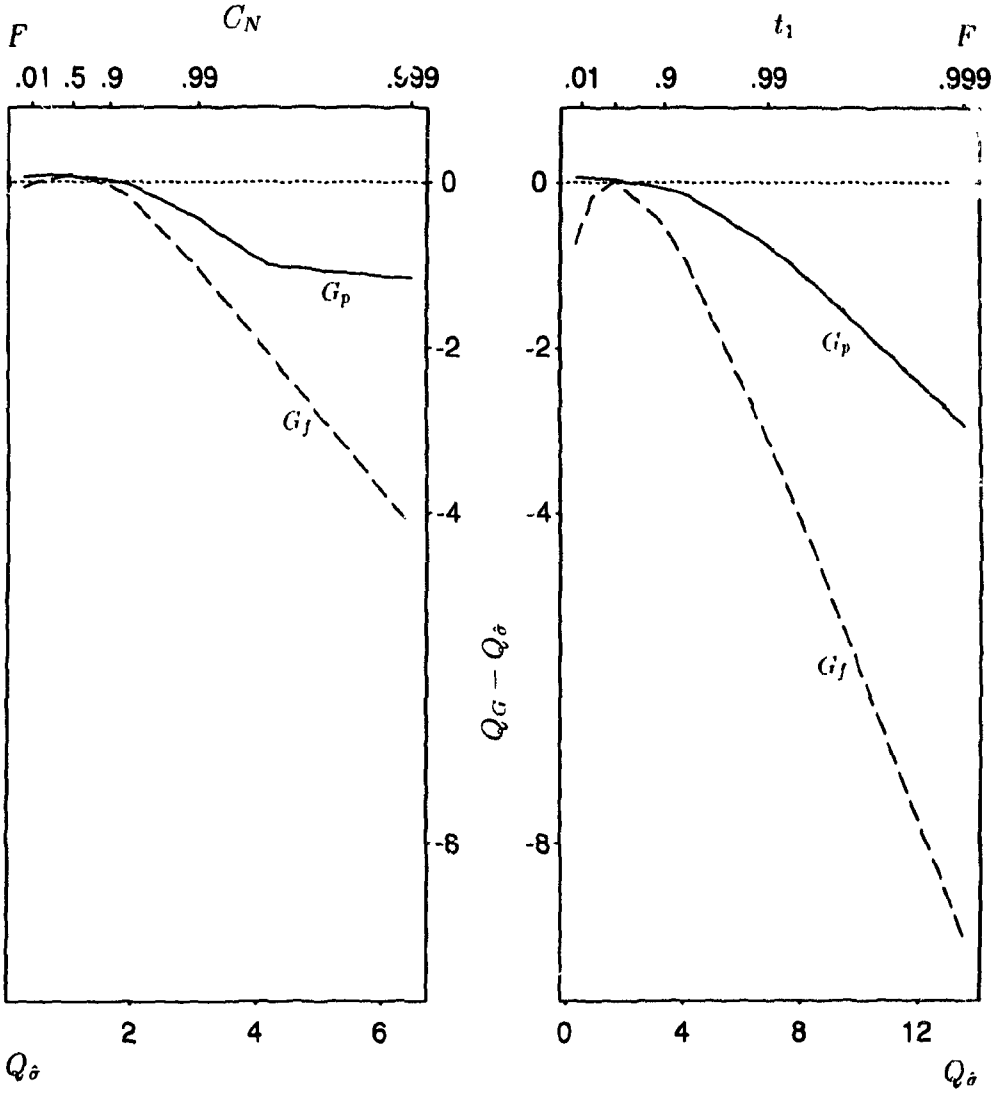


Figure 4.2a
Regression (Huber's): Distributions for $\hat{\rho} = \frac{\hat{\theta}_3 - \theta_{30}}{\gamma\hat{\sigma}}$

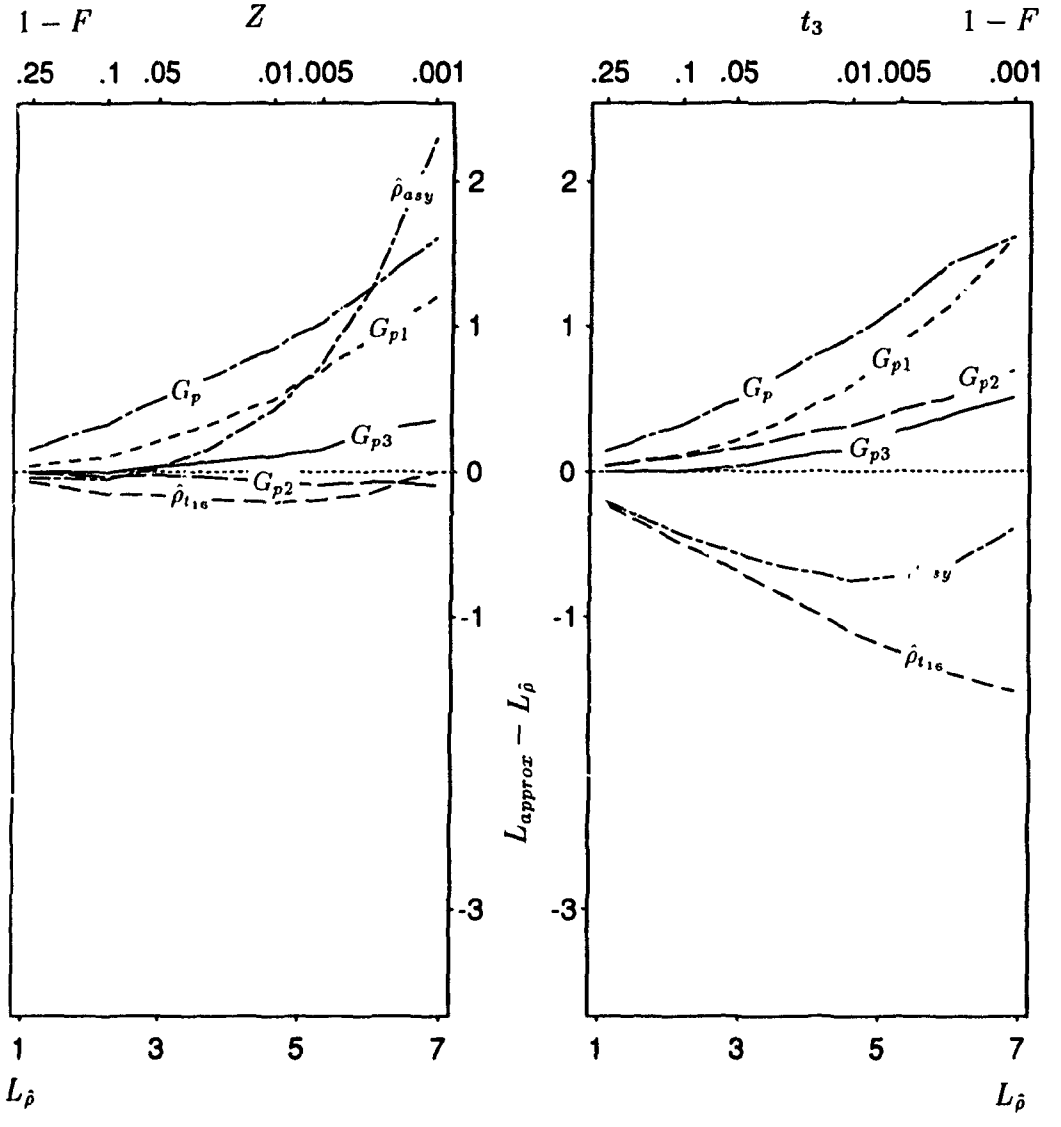
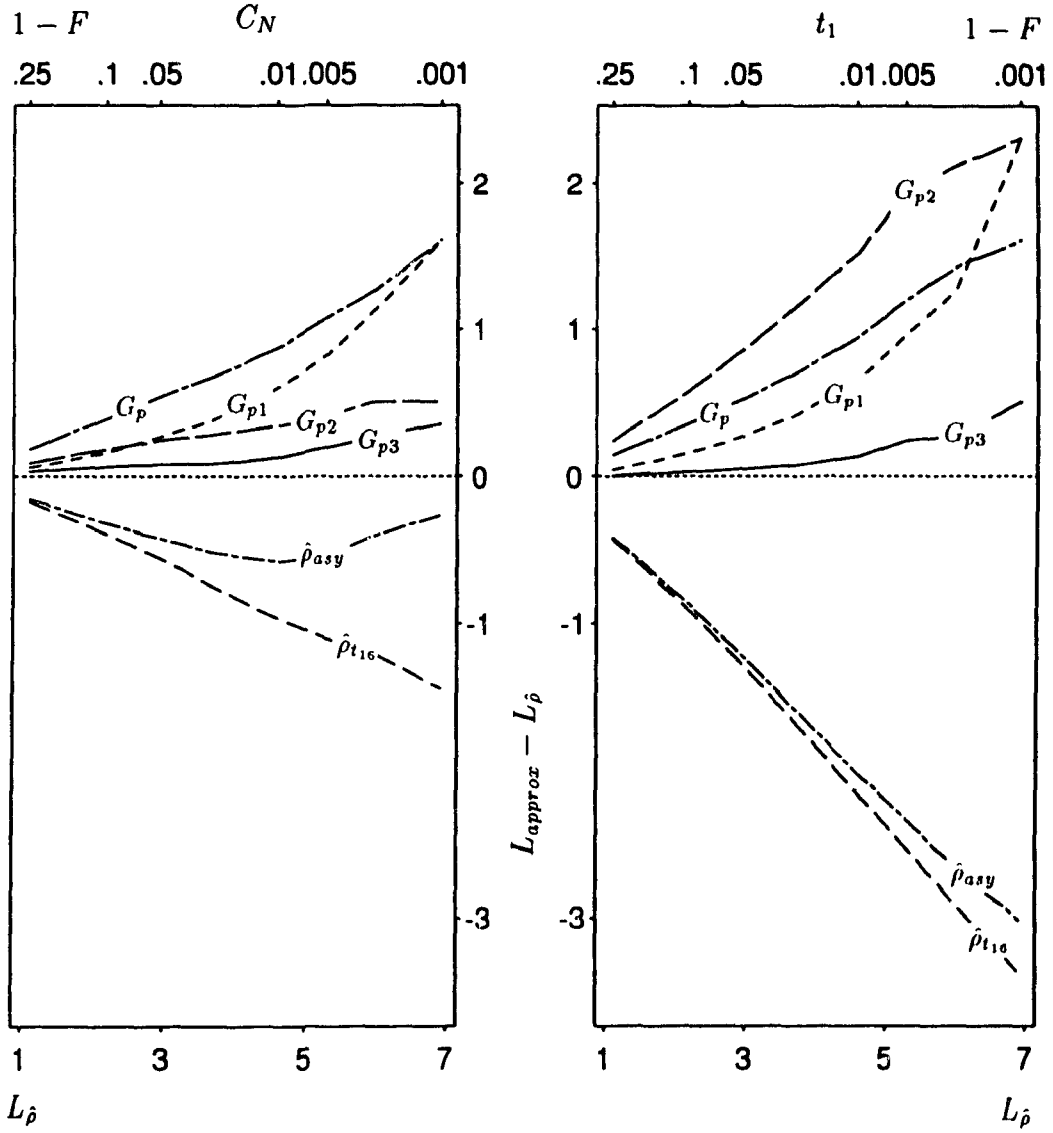


Figure 4.2b
Regression (Huber's): Distributions for $\hat{\rho} = \frac{\hat{\theta}_3 - \theta_{30}}{\gamma\hat{\sigma}}$



# Chapter 5

# Approximation for joint densities

## 5.1 Overview

Up to this point, we have limited our discussion to the approximation for the marginal densities of a single estimator. We have derived $G_p$ for a real-valued function $\rho$ of a multivariate $M$-estimator $\hat{\eta}$ in Chapter 3 and have demonstrated its accuracy in Chapter 4. We now generalize the result to an approximation for the joint densities of a $k$-dimensional real-valued function.

The result that we have obtained is useful for many practical purposes, from studying the random behaviour of an estimator to testing a hypothesis. However, there are still many applications in which an understanding of the joint behaviour of two or more estimators is necessary. For instance, we know that the least squares estimator is a special case of the Huber-type estimator in the multiple regression problem, and the estimators $\hat{\theta}$ and $\hat{\sigma}$ are independent when the underlying distribution is normal. We may want to study the dependence or some conditional properties of the Huber-type estimator in general by taking $c < \infty$ in the Huber's score function $\Psi_c$.

Suppose that we want to compute the joint densities of $k$ components in a $p$-dimensional $M$-estimator $\hat{\eta}$, where $k \leq p$. The problem can be solved, at least theoretically, by any one of the three techniques that we have discussed in Sections 2.3 to 2.5. Recall that the techniques include using the asymptotic distribution, an approximation for the joint density function of $\hat{\eta}$, and a linear approximation for the estimator. We give a brief discussion of these alternatives.

When the asymptotic joint distribution of a multivariate estimator is known, it is definitely the simplest approximation to apply. For the $M$-estimators of practical interest the result is generally available. We have demonstrated in several one-dimensional applications (see Chapter 4) that this alternative gives very good approximation for a symmetrically distributed estimator when the underlying distribution is normal or close to normal. Otherwise, the approximation could be very inaccurate. We expect that the situation is similar for the joint densities in a multi-dimensional problem. Another shortcoming of this approximation is that the finite sample behaviour of an estimator could be arbitrarily far from the asymptotic result. For instance, while the Huber-type estimators $\hat{\theta}$ and $\hat{\sigma}$ are generally dependent in a finite sample problem, they may be asymptotically independent. To obtain a sense of its performance, this approach will be applied in an example in Section 5.3.

The second alternative requires the availability of the joint density function of $\hat{\eta}$ or a good approximation for it. For an $M$-estimator, we have the approximation derived by Field (1982). However, the technique requires us to solve a system of $p$ non-linear equations at each point where the $p$-dimensional density is to be approximated. In addition, solving the problem with $k < p$ may involve a high dimensional integration. Although this may give us a more accurate approximation, the solution becomes

impractical when $p$ is larger than 2 and $k$ is relatively small. Further comments on this approach will be given in Section 5.4.

For the third alternative, we can derive a linear approximation $G$ to each of the $k$ components and use the joint densities of the linear functions as approximations for the true ones. We have examined the performance of a single $G_f$ for approximating marginal densities. It generally improves over the asymptotic result but is still inadequate under some long-tailed distributions. Another concern for the approach is that we need an efficient way to compute accurately the joint distributions of the linear functions. Since the linear approximation $G$ is just a mean, we may write it as the solution of a $k$-dimensional system and apply the result of Field (1982) to approximate the joint densities.

We have shown that our $G_p$ is generally more accurate than the asymptotic result and the linear function $G_f$. In addition, $G_p$ approximates the required densities directly so that we do not need the additional high dimensional integration. This seems to have solved most of the difficulties encountered in the other techniques. It remains to show how the $G_p$ can be generalized to an approximation for the joint densities of $k$ components, where $k > 1$.

In Section 5.2, we extend the approximation $G_p$ for the joint densities of $k$ components in a multi-dimensional $\hat{\eta}$. In fact, we consider a more general problem, that is, we derive an approximation for the joint densities of $k$ real-valued functions of $\hat{\eta}$. The result is applied to two examples in Section 5.3. Section 5.4 gives some general remarks on the generalization. Lastly in Section 5.5, we summarize the numerical results and the plots which are generated in Section 5.3.

## 5.2 Derivation of the approximation

In this section we extend our $G_p$ to an approximation for the joint densities of a $k$-dimensional real-valued function of a $p$-dimensional $M$-estimator $\hat{\eta}$, where $k \leq p$. The notation and assumptions for the derivation in this section are basically the same as those which are defined in Chapter 3. Some minor changes on the notation for the current problem will be stated when they are needed.

Let $\rho(\hat{\eta}) = (\rho_1(\hat{\eta}), \ldots, \rho_k(\hat{\eta}))$ be a real-valued vector. Our objective is to derive an approximation for the joint density of $\rho(\hat{\eta})$ at the point $\rho_0 = (\rho_{10}, \ldots, \rho_{k0})$ under $f$. The development of the approximation parallels that for the marginal density in Section 3.3. We now present the modifications.

To begin, we need the following modified centering lemma.

**Lemma 5.1** *The joint density of $\rho(\hat{\eta})$ at $\rho_0$ under $f$ and that under $h$ are related by*

$$g_f(\rho_0) = \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} E_h\left[exp\left\{-\sum_{j=1}^{p}\alpha_j S_j\right\}\middle|\, \rho(\hat{\eta}) = \rho_0\right] g_h(\rho_0),$$

*where*

$$S = (S_1, \ldots, S_p) = \left\{\sum_{l=1}^{n}\Psi_{jl}(Y_l, t_0)\right\}_{j=1,\ldots,p}.$$

**Proof** Recall from the derivation of Lemma 3.2 that the joint density function of $(S, \hat{\eta})$ under $f$ and that under $h$ are related by

$$g_f(s,t) = \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} exp\left\{-\sum_{j=1}^{p}\alpha_j s_j\right\} g_h(s,t).$$

Integrating both sides of the equality over $\rho(t) = \rho_0$ yields

$$\begin{aligned} g_f(s,\rho_0) &= \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} exp\left\{-\sum_{j=1}^{p}\alpha_j s_j\right\} g_h(s,\rho_0) \\ &= \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} exp\left\{-\sum_{j=1}^{p}\alpha_j s_j\right\} g_h(s|\rho_0) g_h(\rho_0). \end{aligned}$$

Integrating both sides of the last equality with respect to $s$, the result follows.

□

The next step is to simplify the relationship of the two joint density functions by eliminating the conditional expectation. We want to apply similar proportionality arguments as for the one-dimensional case, but there is a problem. Recall that we currently need $2p$ constraints to define the conjugate density $h$, that is, $p$ for the $\alpha$ and $p$ for the $t_0$. To proceed with the arguments, we realize that

$$E_h\left[exp\left\{-\sum_{j=1}^{p}\alpha_j S_j\right\}\middle|\,\rho(\hat{\eta})=\rho_0\right]=1$$

if

$$\sum_{j=1}^{p}\alpha_j S_j \;\propto\; \sum_{r=1}^{k}\{\rho_r(\hat{\eta})-\rho_{r0}\}. \tag{5.1}$$

Note that the particular choice of proportionality in (5.1) is not important, but that this simple choice will illustrate the difficulties in a situation where we need to satisfy just one proportionality.

We use

$$G=(G_1,\ldots,G_k)=\left\{\rho_{r0}+\frac{1}{n}\sum_{l=1}^{n}\sum_{j=1}^{p}\sum_{i=1}^{p}\rho_r^{(i)}(t_0)B_{ij}\Psi_{jl}(Y_l,t_0)\right\}_{r=1,\ldots,k}$$

to approximate $\rho(\hat{\eta})$. From the definition of $S$, the proportionality (5.1) becomes

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\alpha_j\Psi_{jl}(Y_l,t_0)\;\propto\;\sum_{r=1}^{k}\sum_{l=1}^{n}\sum_{j=1}^{p}\sum_{i=1}^{p}\rho_r^{(i)}(t_0)B_{ij}\Psi_{jl}(Y_l,t_0),$$

which is true if

$$\alpha_{j_1}\sum_{r=1}^{k}\sum_{i=1}^{p}\rho_r^{(\cdot)}(t_0)B_{ij_2}=\alpha_{j_2}\sum_{r=1}^{k}\sum_{i=1}^{p}\rho_r^{(i)}(t_0)B_{ij_1}$$

for $1\le j_1,j_2\le p$. This accounts for $p-1$ conditions. Together with the $p+k$ centering conditions

$$E_h\left[\frac{1}{n}\sum_{l=1}^{n}\Psi(Y_l,t_0)\right]=0 \quad \text{and} \quad \rho(t_0)=\rho_0,$$

we have a total of $2p+k-1$ conditions, which exceeds the $2p$ constraints that we need unless $k=1$. Therefore for any $k>1$, we generally cannot find a suitable $h$ which satisfies all the conditions.

To tackle the problem, we have to match the number of constraints that we need and the number of conditions that we have. For our technique, it seems unlikely that we can reduce the number of conditions since the $p+k$ centering conditions are necessary and $p-1$ equalities are needed to satisfy one single proportionality. An alternative is to increase the number of constraints, and possibly the number of conditions for a balance. We proceed as follows.

Our aim is to generate more constraints. However, we need the relationship in Lemma 5.1 and cannot change its basic format. To achieve both objectives, a possibility is to split the $\alpha_j$'s. Writing $\alpha_j=\alpha_{j1}+\cdots+\alpha_{jk}$, $j=1,\ldots,p$, the conditional expectation in Lemma 5.1 now becomes

$$E_h\left[exp\left\{-\sum_{r=1}^{k}\sum_{j=1}^{p}\alpha_{jr}S_j\right\}\middle|\,\rho_1(\hat{\eta})=\rho_{10},\ldots,\rho_k(\hat{\eta})=\rho_{k0}\right]=1,$$

which is trivial if

$$\sum_{j=1}^{p}\alpha_{jr}S_j \;\propto\; \rho_r(\hat{\eta})-\rho_{r0},\quad r=1,\ldots,k.$$

Using $G$ to approximate $\rho(\hat{\eta})$ and the definition of $S$, the last set of proportionalities becomes

$$\sum_{l=1}^{n}\sum_{j=1}^{p}\alpha_{jr}\Psi_{jl}(Y_l,t_0) \;\propto\; \sum_{l=1}^{n}\sum_{j=1}^{p}\sum_{i=1}^{p}\rho_r^{(i)}(t_0)B_{ij}\Psi_{jl}(Y_l,t_0),\quad r=1,\ldots,k,$$

which is true if

$$\alpha_{j_1 r} \sum_{i=1}^{p} \rho_r^{(i)}(t_0) B_{i j_2} = \alpha_{j_2 r} \sum_{i=1}^{p} \rho_r^{(i)}(t_0) B_{i j_1}$$

for $1 \leq j_1, j_2 \leq p$, $r = 1, \ldots, k$. This accounts for $k(p-1)$ conditions. Together with the $p+k$ centering conditions, we have now a total of $(k+1)p$ conditions. This matches the number of constraints that we need, that is, $kp$ for $\alpha_{jr}$, $j = 1, \ldots, p$, $r = 1, \ldots, k$, and $p$ for $t_0$. It follows from Lemma 5.1 that with the conjugate density function $h$ in which the parameters $\alpha$ and $t_0$ are chosen such that the $(k+1)p$ conditions are satisfied, we can approximate the joint density of $\rho(\hat{\eta})$ at $\rho_0$ under $f$ by

$$\left\{ \prod_{l=1}^{n} c_l(t_0) \right\}^{-1} g_h(\rho_0).$$

For the above approximation to be useful, we need to evaluate the joint density $g_h(\rho_0)$. We have derived a linear approximation $G$ for the random vector $\rho(\hat{\eta})$. An obvious choice would be to use $g_{G|h}(\rho_0)$, the density of $G$ at $\rho_0$ under $h$, as an approximation for $g_h(\rho_0)$. Recall that $\rho_0$ is the expected value of $G$ under $h$. Therefore to approximate the joint density $g_{G|h}(\rho_0)$, a one-term Edgeworth approximation gives

$$g_{G|h}(\rho_0) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} + O\left(\frac{1}{n}\right), \tag{5.2}$$

where $|\Sigma|$ is the determinant of the covariance matrix $\Sigma$ of the $k$-dimensional $G$ (McCullagh, 1987, page 150). As for the marginal density approximation, this Edgeworth approximation can be replaced when there exists a better alternative.

Putting the results in this section together, we obtain a multivariate density approximation as follows.

**Theorem 5.1** *Let $\rho(\hat{\eta}) = (\rho_1(\hat{\eta}), \ldots, \rho_k(\hat{\eta}))$ be a real-valued function of a multivariate M-estimator $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_p)$ which solves the system of equations*

$$\frac{1}{n}\sum_{l=1}^{n}\Psi_l(Y_l,\hat{\eta})=0,$$

*where* $k \le p$, $\Psi_l = (\Psi_{1l},\ldots,\Psi_{pl})$, $l = 1,\ldots,n$, *and the* $Y_l$*'s are independent with densities* $f_l(y_l)$. *If the assumptions A1 - A8 in Section 3.2 are satisfied, an approximation for the joint density of* $\rho(\hat{\eta})$ *at* $\rho_0 = (\rho_{10},\ldots,\rho_{k0})$ *under the joint density function* $f = \prod_l f_l$ *is given by*

$$g_p(\rho_0) = \left\{\prod_{l=1}^{n} c_l(t_0)\right\}^{-1} (2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}\,, \tag{5.3}$$

*where*

$$c_l^{-1}(t_0) = \int_{-\infty}^{\infty} exp\left\{\sum_{j=1}^{p}\alpha_j\Psi_{jl}(y_l,t_0)\right\} f_l(y_l)\,dy_l,$$

$|\Sigma|$ *is the determinant of the covariance matrix* $\Sigma$ *of*

$$G = \left\{\rho_r(t_0) + \frac{1}{n}\sum_{l=1}^{n}\sum_{j=1}^{p}\sum_{i=1}^{p}\rho_r^{(i)}(t_0)B_{ij}\Psi_{jl}(Y_l,t_0)\right\}_{r=1,\ ,k}$$

*under the joint conjugate density function* $h = \prod_l h_l$,

$$h_l(y_l) = c_l(t_0)\ exp\left\{\sum_{j=1}^{p}\alpha_j\Psi_{jl}(y_l,t_0)\right\} f_l(y_l),$$

$\alpha_j = \alpha_{j1} + \cdots + \alpha_{jk}$, $j = 1,\ldots,p$ *and* $t_0 = (t_{10},\ldots,t_{p0})$ *are chosen such that*

$$E_h\left[\frac{1}{n}\sum_{l=1}^{n}\Psi_l(Y_l,t_0)\right] = 0,\ \ \rho(t_0) = \rho_0,$$

$$\alpha_{j_1 r}\sum_{i=1}^{p}\rho_r^{(i)}(t_0)B_{ij_2} \;=\; \alpha_{j_2 r}\sum_{i=1}^{p}\rho_r^{(i)}(t_0)B_{ij_1},\ \ 1 \le j_1, j_2 \le p,\ \ r = 1,\ldots,k,$$

*and*

$$B = \{B_{ij}\}_{1\le i,j\le p} = -A^{-1}(t_0),\ \ A(t_0) = E_h\left[\frac{1}{n}\sum_{l=1}^{n}\left.\frac{\partial\Psi_l(Y_l,\eta)}{\partial\eta^T}\right|_{\eta=t_0}\right].$$

□

We define $G_p$ to be the approximation for which the joint density at $\rho_0$ is the normalized $g_p(\rho_0)$ in (5.3), that is,

$$g_{G_p}(\rho_0) = \frac{g_p(\rho_0)}{\int_x g_p(x)\, dx} .$$

In the next section, we will apply the multivariate $G_p$ to some numerical examples. The error and the computational aspect of the approximation will be discussed in Section 5.4. It is clear from their definitions that the marginal density approximation developed in Chapter 3 is simply a special case of the multivariate result with $k = 1$. We will compare the two $G_p$'s numerically in an example in Section 5.3 and will make some general comparison between them in Section 5.4.

## 5.3 Some examples

In this section, we implement the multivariate $G_p$ for two numerical examples which are based on the multiple regression model and the Huber-type estimator defined in Section 2.2. Unless specified otherwise, we adopt the notation and the settings that are defined in Chapter 4.

The main objective of these examples is to demonstrate the accuracy obtained by our $G_p$. Nevertheless, we will compute an asymptotic joint distribution in the first example and will implement an adjustment to the $G_p$ in the second one for comparison. The performance of the approximations is examined under the error distributions $Z$ and $t_3$.

The joint distributions of $G_p$ are computed by numerical integration of the joint density approximation given in (5.3), and the joint distributions of the multivariate estimator $\hat{\rho} = \rho(\hat{\eta})$ are based on 100,000 simulations. In addition to the computer subroutines that are mentioned in Chapter 4, we also need the subroutine D01DAF in NAG for multi-dimensional integrations. Numerical results and some contour plots are summarized in Section 5.5.

Example 1:

Our first example examines the joint behaviour of the two-dimensional estimator $\hat{\rho} = (\hat{\theta}_3, \hat{\theta}_4)$, that is, $k = 2$. We have experienced the good performance of both the asymptotic approximation and the $G_p$ for the marginal densities of $\hat{\theta}_3$. We hope that their performance is similar for joint density approximation. In addition, we expect that the approximations would perform the best under normal distribution and when the scale parameter is known. For this reason, we assume that $\sigma_0 = 1$ is known in the

estimation of this example. Therefore the multiple regression model and the score functions for the estimator simplify to

$$Y_l = X_l^T \theta + \varepsilon_l, \quad l = 1, \ldots, n,$$

and

$$\Psi_{jl}(Y_l, \eta) = \Psi_c(Y_l - X_l^T \theta) X_{lj}, \quad j = 1, \ldots, p,$$

respectively. Note that we now have $p : n = 4 : 20$.

In the simulation, we take $\theta_0 = 0$ for simplicity. The asymptotic joint distribution of $\hat{\rho}$ from (2.5) is given by

$$\hat{\rho}_{asy} \sim N\left(0, \frac{E_f[\Psi_c^2(\varepsilon_1)]}{\{E_f[I_c(\varepsilon_1)]\}^2}(X^T X)_\rho^{-1}\right),$$

where $I_c(x)$ equals 1 if $|x| < c$, and 0 otherwise, $(X^T X)_\rho^{-1}$ is the lower-right corner of $(X^T X)^{-1}$ of order $2 \times 2$, and

| $f_\varepsilon$ | $Z$ | | $t_3$ | |
|---|---|---|---|---|
| $\frac{E_f[\Psi_c^2(\varepsilon_1)]}{\{E_f[I_c(\varepsilon_1)]\}^2}(X^T X)_\rho^{-1}$ | .6435 | −.0094 | .9516 | −.0139 |
| | −.0094 | .5363 | −.0139 | .7930 |

For the linear approximation $G$, we have

$$A = -\frac{1}{n} E_h[X^T D X] \quad \text{and} \quad B = n\left\{X^T E_h[D] X\right\}^{-1},$$

where $D$ is an $n$ by $n$ diagonal matrix with diagonal elements $I_c(r_l)$, $r_l = Y_l - X_l^T t_0$, $l = 1, \ldots, n$. Therefore

$$G = (G_3, G_4) = \left\{ t_{i0} + \sum_{l=1}^{n} \sum_{j=1}^{p} \left\{X^T E_h[D] X\right\}_{ij}^{-1} \Psi_c(r_l) X_{lj} \right\}_{i=3,4},$$

and the one-term Edgeworth approximation (5.2) of its joint density at the expected value is given by

$$g_{G|h}(t_{30}, t_{40}) = \frac{1}{2\pi\sqrt{\sigma^2_{G_3|h}\sigma^2_{G_4|h} - \sigma^2_{G_3G_4|h}}}.$$

where $\sigma^2_{G_3|h}$ and $\sigma^2_{G_4|h}$ are the variances of $G_3$ and $G_4$ under $h$ respectively, and $\sigma_{G_3G_4|h}$ is the covariance of $G_3$ and $G_4$ under $h$.

Numerical results are given in Tables 5.1a and 5.1b. As we expected, the situation is very similar to that of the one-dimensional problems. In particular, both $\hat{\rho}_{asy}$ and $G_p$ generate excellent approximations under $Z$. $G_p$ seems to be slightly better than $\hat{\rho}_{asy}$, but the improvement is hardly significant. For the case under $t_3$, it is clear that $G_p$ provides very good results for both marginal distributions.

To obtain a better picture of the overall performance of the approximations, Figures 5.1a and 5.1b plot two contour maps based on the results in Tables 5.1a and 5.1b respectively. In brief, $\hat{\rho}$, $\hat{\rho}_{asy}$ and $G_p$ show almost no difference under $Z$, and both approximations $\hat{\rho}_{asy}$ and $G_p$ give very good results around the center under $t_3$. Moreover, $G_p$ gives consistently good approximation over the entire region.

Example 2:

We have encountered some problems in the approximation for the marginal densities of $\hat{\sigma}$. We now apply the multivariate $G_p$ for the joint densities of $\hat{\rho} = (\hat{\theta}_3, \hat{\sigma})$. The setting is exactly the same as those are defined in Sections 2.2 and 4.3. A derivation of the linear function $G$ and other details for the current approximation can be found in the two sections. In addition to the basic $G_p$, we also evaluate $G_{p2}$ to obtain a sense of adjustment. Recall that $G_{p2}$ is basically the same as $G_p$ except that a constant adjustment is applied in the computation of $G$. Details of the adjustment are given

in Appendix A.

Numerical results are summarized in Tables 5.2a and 5.2b and are plotted in Figures 5.2a and 5.2b. It is clear from the two tables that the multivariate $G_p$ performs basically the same as the univariate $G_p$ does for the marginal distributions of $\hat{\theta}_3$ and $\hat{\sigma}$, except that the former one is now doing both jobs simultaneously. It gives very good approximations for the marginal distributions of $\hat{\theta}_3$ and at the same time suffers a similar problem for the approximation of $\hat{\sigma}$.

We observe from the plots some general performance of the approximations. In addition to the marginal behaviour, the plots show that $G_p$ has provided a fair approximation to the shape of the joint distributions. However, the approximation is shifted to the right in the direction of $\hat{\sigma}$. This causes the conditional approximation for $\hat{\theta}_3$ to become very inaccurate when $\hat{\sigma}$ is small.

Based on this observation, a constant adjustment in the centering procedure seems adequate for an improvement. For this reason, we compute the adjusted approximation $G_{p2}$. It is clear from the results that the constant adjustment improves significantly from the basic $G_p$ over the entire domain. Our objective is not trying to find the best adjustment for this example. However, $G_{p2}$ is clearly good enough to illustrate the effect of an adjustment.

Lastly, we demonstrate another possible application of the joint density approximation. In Section 4.3, we use the univariate $G_p$ to approximate the marginal densities of a studentized $t$-ratio. With the joint density approximation of $\hat{\theta}_3$ and $\hat{\sigma}$, we can evaluate the marginal densities of the ratio by numerical integration. We compute the marginal approximations by using the multivariate $G_p$ and $G_{p2}$ and summarize the results into Table 5.3. For comparison, we restate their counterparts by using the

univariate $G_p$ and $G_{p2}$ in the same table.

Generally, the approximation from the multivariate $G_p$ improves slightly over that from the univariate $G_p$. The improvement possibly comes from the fact that we do not need to approximate the non-linear ratio by a linear function. When a constant adjustment is applied, the approximations are clearly improved with both univariate and multivariate approaches. However, the degrees of improvement are slightly different. While the multivariate approach seems to be better than the univariate one under $t_3$, the situation reverses under $Z$.

## 5.4 Discussion

In this chapter, we have derived an approximation for the joint densities of a multivariate function of an $M$-estimator. The approximation is an extension of the marginal density approximation developed in Chapter 3. The multivariate approximation is applied to several examples. We now give some general comments on the results.

With this extension, we can study the joint and the conditional behaviour of a multivariate estimator. In addition, since the multivariate $G_p$ is developed under the same assumptions as for the univariate one, we can always apply it to provide an alternative for the marginal densities of an estimator. We have demonstrated such a possibility in an example. When no adjustment is applied, it still improves the univariate approximation by eliminating the error induced in the linearization of a non-linear function. For a more complicated function, this alternative would be proven more beneficial.

When we compare the performance of the multivariate $G_p$ and the univariate $G_p$, in particular on the marginal distribution approximation, we can see that the two approaches generally possess very similar characteristics. This is not unexpected since we have used basically the same arguments to derive both $G_p$'s. In fact, we can expect that the error of both approximations are of the same order.

Consider the development of the two $G_p$'s. We use the linear function $G$ for an approximation of the density at the expected value of $\hat{\rho}$ under $h$, apply $G$ and similar proportionality arguments to simplify the conditional expectation, and use a one-term Edgeworth approximation for the density at the expected value of $G$. All these give the same order of error on both $G_p$'s. A major difference in the derivations is that

we have $k$ proportionalities instead of one for the multivariate $G_p$, but this does not affect the order of error.

Concerning the computational effort, we see that at each point the joint density is to be approximated, a system of $(k+1)p$ non-linear equations needs to be solved. This may cause some problems when $k$ and $p$ are both large. Comparatively, the joint density approximation derived by Field (1982) would become more attractive when $k$ is close to $p$ and $p$ is large, in which case a $p$-dimensional non-linear system needs to be solved at each point the $p$-dimensional joint density approximation is computed, and a $(p-k)$-fold numerical integration is needed. Our $G_p$ does not require the last integration. In addition, we have another advantage that we do not need to worry about the region of integration. If we use the joint density approximation $g_T(t_0)$ of Field (1982) (see Section 2.4) to approximate $g_\rho(\rho_0)$, we need to integrate $g_T(t_0)$ over the region $\rho(t_0) = \rho_0$. Unless the region can be expressed in a closed form, we need to determine it numerically.

## 5.5 Numerical results

Table 5.1a
Regression (Huber's with a known scale): Joint distributions for $\hat{\rho} = (\hat{\theta}_3, \hat{\theta}_4)$

| | | $Z$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\theta}_4$ | $\hat{\theta}_3$ | | | | | | | |
| | | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 |
| $\hat{\rho}$ | -3 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| | -2 | .0000 | .0000 | .0003 | .0014 | .0026 | .0029 | .0030 | .0030 |
| | -1 | .0000 | .0004 | .0083 | .0417 | .0755 | .0840 | .0846 | .0846 |
| | 0 | .0000 | .0029 | .0513 | .2464 | .4463 | .4958 | .4990 | .4990 |
| | 1 | .0001 | .0056 | .0951 | .4556 | .8205 | .9098 | .9154 | .9154 |
| | 2 | .0001 | .0062 | .1046 | .4977 | .8940 | .9911 | .9972 | .9973 |
| | 3 | .0001 | .0063 | .1049 | .4990 | .8963 | .9938 | .9999 | 1.0000 |
| $\hat{\rho}_{asy}$ | -3 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| | -2 | .0000 | .0000 | .0003 | .0015 | .0028 | .0031 | .0032 | .0032 |
| | -1 | .0000 | .0005 | .0087 | .0420 | .0764 | .0855 | .0860 | .0860 |
| | 0 | .0000 | .0031 | .0520 | .2475 | .4457 | .4967 | .5000 | .5000 |
| | 1 | .0001 | .0057 | .0967 | .4560 | .8164 | .9082 | .9139 | .9140 |
| | 2 | .0001 | .0063 | .1059 | .4984 | .8909 | .9906 | .9968 | .9969 |
| | 3 | .0001 | .0063 | .1063 | .5000 | .8938 | .9937 | .9999 | 1.0000 |
| $G_p$ | -3 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| | -2 | .0000 | .0000 | .0003 | .0015 | .0027 | .0031 | .0031 | .0031 |
| | -1 | .0000 | .0005 | .0086 | .0416 | .0758 | .0847 | .0853 | .0853 |
| | 0 | .0000 | .0030 | .0516 | .2474 | .4461 | .4968 | .5000 | .5000 |
| | 1 | .0001 | .0056 | .0960 | .4564 | .8178 | .9090 | .9147 | .9147 |
| | 2 | .0001 | .0062 | .1051 | .4984 | .8917 | .9907 | .9968 | .9969 |
| | 3 | .0001 | .0062 | .1055 | .5000 | .8945 | .9938 | .9999 | 1.0000 |

Table 5.1b
Regression (Huber's with a known scale): Joint distributions for $\hat{\rho} = (\hat{\theta}_3, \hat{\theta}_4)$

| | | $t_3$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\theta}_4$ | $\hat{\theta}_3$ | | | | | | | |
| | | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 |
| $\hat{\rho}$ | -3 | .0001 | .0002 | .0004 | .0008 | .0012 | .0014 | .0014 | .0014 |
| | -2 | .0002 | .0011 | .0035 | .0092 | .0150 | .0179 | .0188 | .0189 |
| | -1 | .0006 | .0045 | .0233 | .0690 | .1151 | .1350 | .1393 | .1399 |
| | 0 | .0015 | .0135 | .0790 | .2486 | .4172 | .4836 | .4966 | .4981 |
| | 1 | .0023 | .0231 | .1358 | .4289 | .7215 | .8339 | .8545 | .8569 |
| | 2 | .0028 | .0273 | .1579 | .4919 | .8241 | .9539 | .9780 | .9808 |
| | 3 | .0030 | .0282 | .1615 | .5006 | .8382 | .9705 | .9953 | .9982 |
| | 4 | .0030 | .0284 | .1618 | .5013 | .8393 | .9718 | .9967 | .9997 |
| $\hat{\rho}_{asy}$ | -3 | .0000 | .0000 | .0001 | .0002 | .0003 | .0004 | .0004 | .0004 |
| | -2 | .0000 | .0002 | .0018 | .0060 | .0103 | .0121 | .0123 | .0124 |
| | -1 | .0001 | .0025 | .0192 | .0640 | .1100 | .1279 | .1306 | .1307 |
| | 0 | .0005 | .0098 | .0748 | .2475 | .4222 | .4896 | .4995 | .5000 |
| | 1 | .0009 | .0174 | .1319 | .4333 | .7358 | .8516 | .8684 | .8693 |
| | 2 | .0010 | .0199 | .1506 | .4936 | .8368 | .9678 | .9867 | .9877 |
| | 3 | .0011 | .0202 | .1526 | .4998 | .8471 | .9795 | .9986 | .9997 |
| | 4 | .0011 | .0202 | .1527 | .5000 | .8474 | .9799 | .9990 | 1.0000 |
| $G_p$ | -3 | .0000 | .0001 | .0003 | .0006 | .0009 | .0011 | .0012 | .0012 |
| | -2 | .0002 | .0008 | .0032 | .0081 | .0131 | .0158 | .0165 | .0167 |
| | -1 | .0005 | .0042 | .0224 | .0663 | .1115 | .1309 | .1348 | .1353 |
| | 0 | .0012 | .0123 | .0771 | .2475 | .4203 | .4874 | .4988 | .4999 |
| | 1 | .0019 | .0205 | .1329 | .4310 | .7303 | .8440 | .8627 | .8645 |
| | 2 | .0023 | .0240 | .1532 | .4914 | .8298 | .9592 | .9810 | .9832 |
| | 3 | .0024 | .0248 | .1564 | .4994 | .8423 | .9740 | .9963 | .9986 |
| | 4 | .0025 | .0249 | .1567 | .5000 | .8432 | .9750 | .9975 | .9997 |

Table 5.2a
Regression (Huber's): Joint distributions for $\hat{\rho} = (\hat{\theta}_3, \hat{\sigma})$

| | | $Z$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\sigma}$ | $\hat{\theta}_3$ | | | | | | | |
| | | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 |
| $\hat{\rho}$ | 0.5 | .0000 | .0000 | .0003 | .0014 | .0023 | .0026 | .0026 | .0026 |
| | 1.0 | .0000 | .0025 | .0448 | .2116 | .3803 | .4224 | .4249 | .4249 |
| | 1.5 | .0001 | .0057 | .1015 | .4885 | .8777 | .9718 | .9776 | .9777 |
| | 2.0 | .0001 | .0059 | .1039 | .4996 | .8975 | .9941 | .9999 | 1.0000 |
| | 2.5 | .0001 | .0059 | .1039 | .4996 | .8975 | .9941 | .9999 | 1.0000 |
| $G_p$ | 0.5 | .0000 | .0000 | .0000 | .0001 | .0002 | .0002 | .0002 | .0002 |
| | 1.0 | .0000 | .0014 | .0237 | .1133 | .2030 | .2253 | .2266 | .2267 |
| | 1.5 | .0001 | .0054 | .0968 | .4681 | .8393 | .9307 | .9361 | .9362 |
| | 2.0 | .0001 | .0058 | .1033 | .4999 | .8965 | .9940 | .9998 | .9998 |
| | 2.5 | .0001 | .0058 | .1034 | .5000 | .8966 | .9942 | .9999 | 1.0000 |
| $G_{p2}$ | 0.5 | .0000 | .0000 | .0005 | .0023 | .0041 | .0046 | .0046 | .0046 |
| | 1.0 | .0000 | .0028 | .0499 | .2399 | .4299 | .4770 | .4798 | .4799 |
| | 1.5 | .0001 | .0057 | .1018 | .4921 | .8825 | .9785 | .9842 | .9843 |
| | 2.0 | .0001 | .0058 | .1034 | .5000 | .8966 | .9942 | .9999 | 1.0000 |
| | 2.5 | .0001 | .0058 | .1034 | .5000 | .8966 | .9942 | .9999 | 1.0000 |

Table 5.2b
Regression (Huber's): Joint distributions for $\hat{\rho} = (\hat{\theta}_3, \hat{\sigma})$

| | | $t_3$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\sigma}$ | $\hat{\theta}_3$ | | | | | | | |
| | | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 |
| $\hat{\rho}$ | 0.5 | .0000 | .0000 | .0001 | .0003 | .0004 | .0005 | .0005 | .0005 |
| | 1.0 | .0001 | .0022 | .0193 | .0746 | .1289 | .1462 | .1481 | .1482 |
| | 1.5 | .0013 | .0156 | .1067 | .3488 | .5914 | .6804 | .6947 | .6960 |
| | 2.0 | .0030 | .0280 | .1582 | .4779 | .7980 | .9258 | .9506 | .9537 |
| | 2.5 | .0037 | .0316 | .1686 | .4986 | .8293 | .9638 | .9911 | .9947 |
| | 3.0 | .0040 | .0322 | .1700 | .5009 | .8325 | .9676 | .9953 | .9991 |
| | 3.5 | .0040 | .0323 | .1701 | .5011 | .8329 | .9680 | .9958 | .9996 |
| $G_p$ | 0.5 | .0000 | .0000 | .0000 | .0000 | .0001 | .0001 | .0001 | .0001 |
| | 1.0 | .0000 | .0007 | .0083 | .0353 | .0623 | .0699 | .0706 | .0706 |
| | 1.5 | .0006 | .0104 | .0825 | .2890 | .4955 | .5676 | .5773 | .5779 |
| | 2.0 | .0021 | .0234 | .1469 | .4637 | .7804 | .9039 | .9252 | .9272 |
| | 2.5 | .0030 | .0279 | .1619 | .4963 | .8308 | .9647 | .9897 | .9924 |
| | 3.0 | .0032 | .0287 | .1637 | .4997 | .8357 | .9707 | .9961 | .9991 |
| | 3.5 | .0033 | .0288 | .1639 | .5000 | .8361 | .9712 | .9967 | .9996 |
| $G_{p2}$ | 0.5 | .0000 | .0000 | .0001 | .0003 | .0006 | .0006 | .0006 | .0006 |
| | 1.0 | .0001 | .0017 | .0180 | .0731 | .1282 | .1446 | .1462 | .1462 |
| | 1.5 | .0009 | .0138 | .1023 | .3471 | .5918 | .6803 | .6932 | .6941 |
| | 2.0 | .0024 | .0251 | .1530 | .4776 | .8022 | .9300 | .9527 | .9550 |
| | 2.5 | .0031 | .0283 | .1627 | .4979 | .8331 | .9675 | .9927 | .9955 |
| | 3.0 | .0032 | .0287 | .1638 | .4998 | .8359 | .9709 | .9964 | .9993 |
| | 3.5 | .0033 | .0288 | .1639 | .5000 | .8361 | .9712 | .9967 | .9997 |

Table 5.3
Regression (Huber's): Tail probabilities for $\hat{\rho} = \frac{\hat{\theta}_3 - \theta_{30}}{\gamma\hat{\sigma}}$

| | | (3.11) | | (5.3) | |
|---|---|---|---|---|---|
| $1-F$ | std. err. | $G_p$ | $G_{p2}$ | $G_p$ | $G_{p2}$ |
| | | $Z$ | | | |
| .25 | .0014 | .2225 | .2518 | .2253 | .2550 |
| .1 | .0009 | .0745 | .1032 | .0773 | .1069 |
| .05 | .0007 | .0312 | .0513 | .0331 | .0543 |
| .025 | .0005 | .0133 | .0260 | .0145 | .0282 |
| .01 | .0003 | .0043 | .0108 | .0049 | .0121 |
| .005 | .0002 | .0018 | .0055 | .0021 | .0064 |
| .0025 | .0002 | .0007 | .0027 | .0009 | .0032 |
| .001 | .0001 | .0002 | .0011 | .0003 | .0014 |
| | | $t_3$ | | | |
| .25 | .0014 | .2239 | .2419 | .2282 | .2463 |
| .1 | .0009 | .0743 | .0915 | .0782 | .0961 |
| .05 | .0007 | .0313 | .0431 | .0339 | .0465 |
| .025 | .0005 | .0127 | .0199 | .0143 | .0220 |
| .01 | .0003 | .0040 | .0073 | .0047 | .0085 |
| .005 | .0002 | .0016 | .0033 | .0019 | .0039 |
| .0025 | .0002 | .0006 | .0015 | .0008 | .0019 |
| .001 | .0001 | .0002 | .0005 | .0002 | .0007 |

Figure 5.1a
Regression (Huber's with a known scale): Contour plots for $\hat{\rho} = (\hat{\theta}_3, \hat{\theta}_4)$

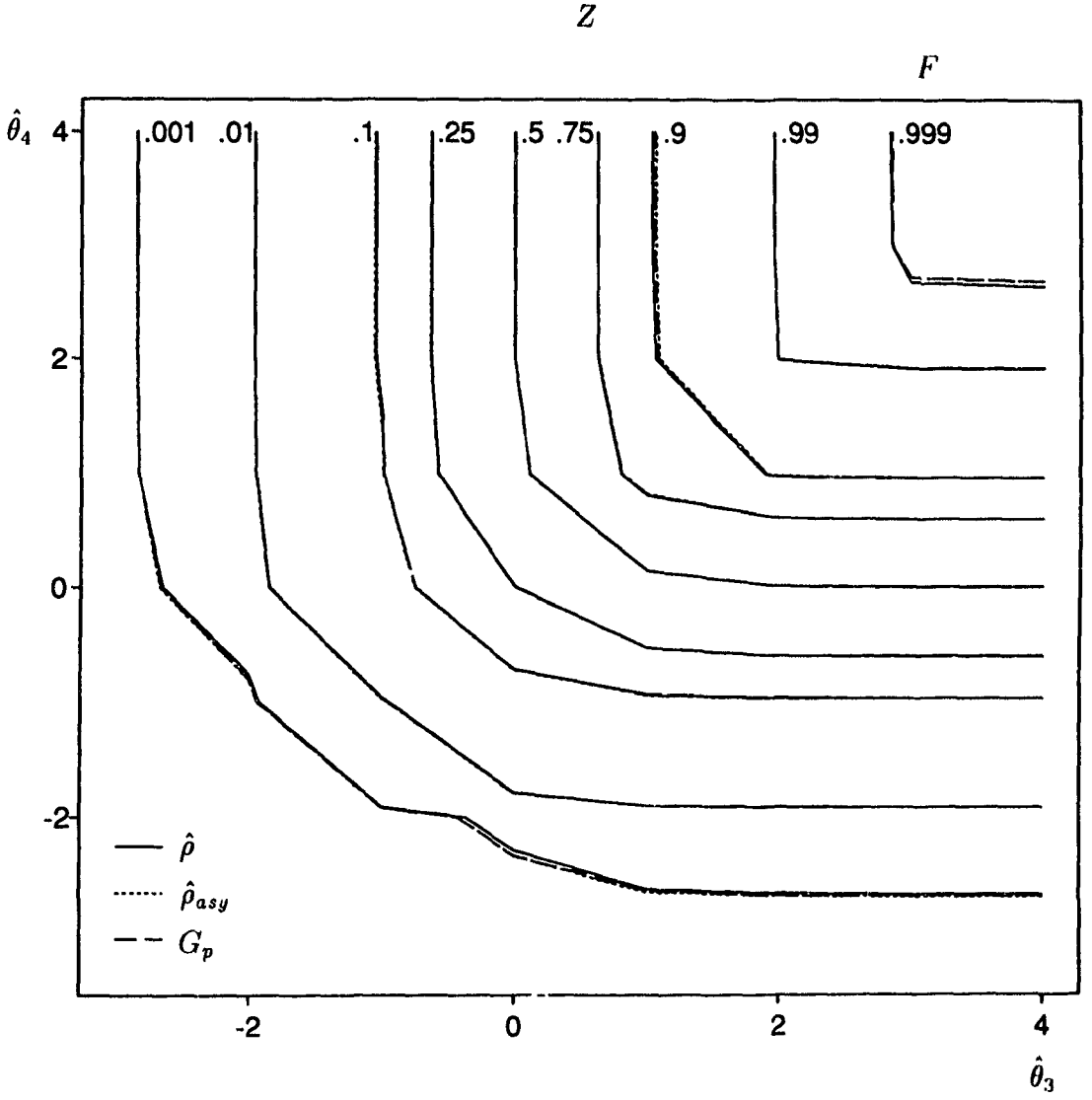


Figure 5.1b
Regression (Huber's with a known scale): Contour plots for $\hat{\rho} = (\hat{\theta}_3, \theta_4)$

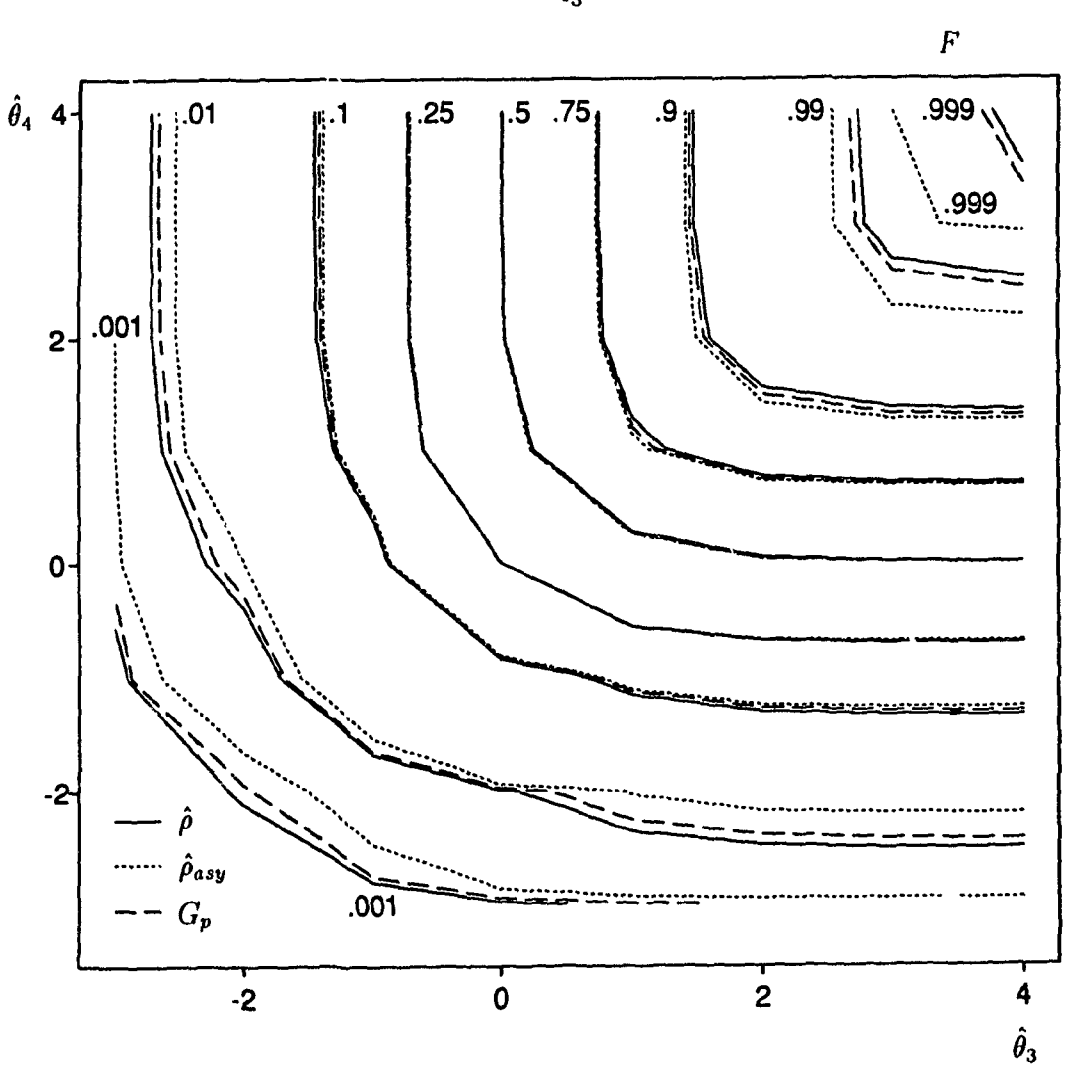


Figure 5.2a
Regression (Huber's): Contour plots for $\hat{\rho} = (\hat{\theta}_3, \hat{\sigma})$

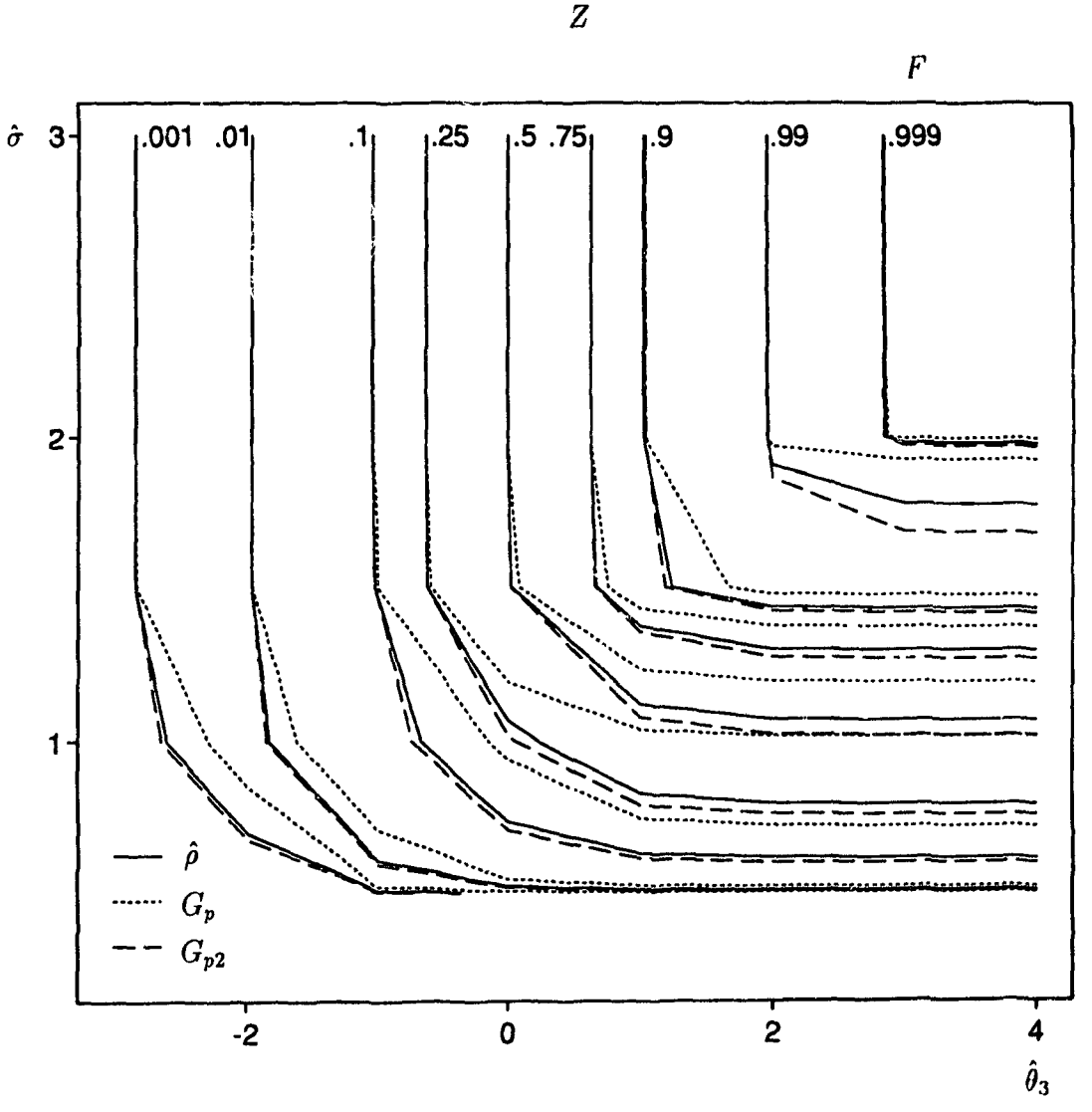
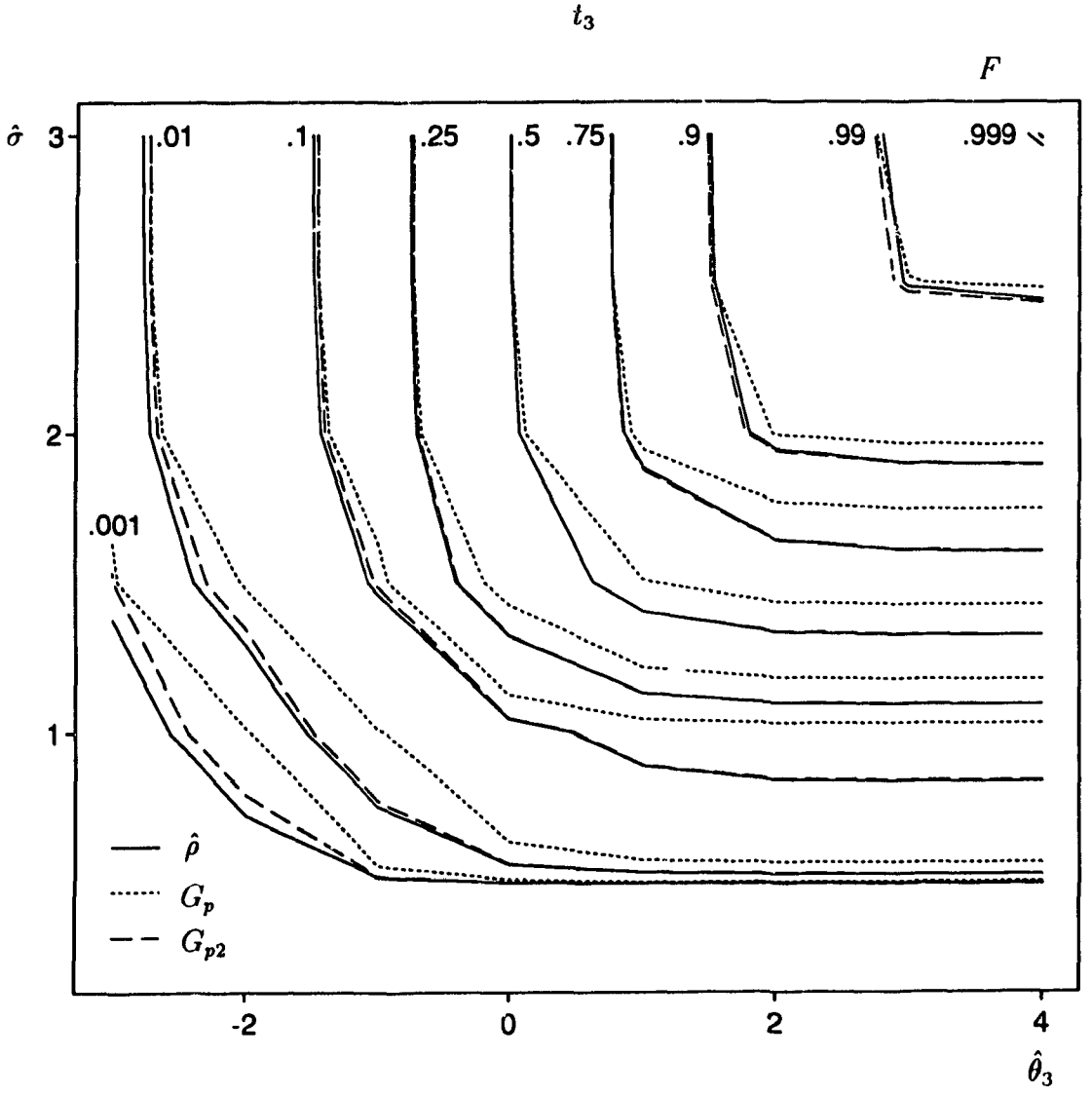


Figure 5.2b
Regression (Huber's): Contour plots for $\hat{\rho} = (\hat{\theta}_3, \hat{\sigma})$



# Chapter 6

# Conclusion

## 6.1 Summary

In this thesis, we have developed a technique to approximate the marginal densities of a multivariate $M$-estimator $\hat{\eta}$, the solution of a non-linear system

$$\frac{1}{n}\sum_{l=1}^{n}\Psi_{jl}(Y_l,\hat{\eta}) = 0, \quad j = 1,\ldots,p,$$

where $Y_l$'s are independent $m$-dimensional random observations from the densities $f_l$'s involving an unknown $p$-dimensional parameter $\eta$. The general problem and some background information are given in Chapter 1.

Under some regularity conditions, our primary result for a real-valued function of $\hat{\eta}$ is derived in Chapter 3. We then generalize the result and derive a joint density approximation for a $k$-dimensional function of the estimator in Chapter 5. The basic idea of our approach can be summarized as follows.

To approximate the density of the function $\hat{\rho} = \rho(\hat{\eta})$ at $\rho_0$ under $f$, we first recenter $f$ to a conjugate density $h$ such that $E_h[\hat{\rho}] = \rho_0$, then approximate the density of $\hat{\rho}$ at the expected value under $h$, and finally transform the approximation under $h$ to

an approximation under $f$.

We implement the approximation for several examples in Chapters 4 and 5. The numerical results show that the approximation is generally accurate over a wide range of underlying distributions, from normal to Cauchy. Nevertheless, some adjustments seem necessary when a scale estimator is part of the statistic for which the density is to be approximated. We propose several adjustments in Chapter 3 and obtain satisfactory improvements in the approximations from them.

We review some existing techniques including the work of Field (1982), Tingley and Field (1990), and DiCiccio and Martin (1991) for our problem in Chapter 2 and try to relate the different approaches in Chapter 3. Specifically, we discuss the close relationship among our approach and the approaches of Field, and Tingley and Field, and establish some formal connections between our result and the work of DiCiccio and Martin.

A practical issue for the different approximations concerns their computational requirements. Basically, the linear approximation for $\hat{\rho}$ derived by Tingley and Field is the simplest and requires the least computational effort, but it generally becomes inadequate in the tails. On the other hand, the joint density approximation developed by Field provides very accurate approximations but it requires substantial computation. Our technique can be viewed as a solution to balance the accuracy and the computational requirement.

Consider a general $k$-dimensional function of the estimator. For the estimator approach, once a linear approximation is derived for each of the components of the $k$-vector, the major effort goes into the evaluation of the joint distribution of the linear functions. The approximation by Field computes the joint densities of $\hat{\eta}$ and

requires additional integrations to obtain the $k$-dimensional densities, and at each point where the joint density is needed, a $p$-dimensional non-linear system must be solved. On the other hand, our technique approximates the $k$-dimensional densities so that no additional integration is required. However, at each point the approximation is computed, a $(k+1)p$-dimensional non-linear system must be solved. It is clear that our approach is particularly useful when $p$ is large and $k$ is small.

Another aspect of our density approximation is about its error. We obtain some results in Chapter 3 but fail to derive the order of the error in general. In particular, we show that the approximation is exact for the least squares estimators of the regressors. Although we may not need an approximation for a least squares estimator, a good approximation should possess this property.

## 6.2 Concluding remarks

For our density approximation, two questions are not completely answered. The first one is about the overall rate of error. A general solution to this problem is difficult to obtain. An answer possibly requires additional assumptions on the density function of the estimator. The second question is about the adjustment. Although we suggest some approaches to the problem, it remains unclear which approach is the best in individual cases. We consider these two questions to be fundamentally important and are still working on the answers.

We demonstrate that the approximation behaves consistently in a simple Mallows problem. A large scale numerical study will be conducted to understand the behaviour of the approximation over different classes of estimators and different underlying distributions. In addition, we have discussed some related techniques in the thesis. Numerical comparisons would be helpful for understanding their relative performance. Note that the computer programs that we write in the development are quite general and their efficiency is not our main concern. Two sample programs are included in Appendix C. For any practical purposes and large scale computation, some tailor-made programs may be needed.

Besides using the approximation to study the behaviour of an estimator, another application is to use it for statistical inferences. We illustrate the possibility through a simple testing problem. However, the true value of the parameter is generally unknown and a location-scale invariant test statistic may not be available. In those situations, the idea of Tingley and Field (1990) may be useful. In brief, when $\hat{\rho}$ is used as a test statistic, the density $g_{f|\eta_0}$ of $\hat{\rho}$ is first approximated by $g_{f|\hat{\eta}_{obs}}$, where $\hat{\eta}_{obs}$ is the observed value of $\hat{\eta}$. An exponential tilt is then applied to force $g_{f|\hat{\eta}_{obs}}$

to satisfy the hypothesis under testing. The details of this possibility are still being worked out.

We aim to approximate the marginal densities of the $M$-estimators and believe that we have found a partial solution to the problem. However, the examples that we use are limited to the functions which are asymptotically normal. We attempt to apply the approximation to a situation where the asymptotic distribution is non-normal. Specifically, we tried to implement our technique to the $\tau$-test proposed by Ronchetti (see Hampel et al., 1986, Chapter 7) and have not yet obtained a satisfactory approximation. Details of the problem is given in Appendix B. It would be interesting to see if some modifications can be applied to make our technique useful in the situation.

Finally, it is pointed out that the density of an $M$-estimator may not exist in some situations. Our objective is not to justify when it will exist. However, even if it does not, the density of the linear approximation $G$ at the expected value can exist and may be used to approximate the distribution of the estimator. This seems promising but careful investigation of the behaviour of the approximation in those situations is needed.

# Appendix A

# Computation of adjustments

## A.1 General remarks

In Section 3.5 we propose three adjustments to improve our density approximation and discuss the motivation of the adjustments. The central idea of the proposals is to replace the linear approximation

$$G = \rho(t_0) + \frac{1}{n}\sum_{l=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{p}\Psi_{jl}(Y_l, t_0)B_{kj}(t_0)\frac{\partial\rho(t_0)}{\partial\eta_k}$$

by

$$G = \rho(t_\mu) + \frac{1}{n}\sum_{l=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{p}\Psi_{jl}(Y_l, t_0)B_{kj}(t_0)\frac{\partial\rho(t_\mu)}{\partial\eta_k},$$

where $t_\mu$ incorporates adjustments to $t_0$ related to the expected value of the error term

$$e_T = \left\{\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{j_3=1}^{p}(\overline{\Psi}_{j_1} - \mu_{j_1})(\overline{\Psi}_{j_2}^{(j_3)} - \mu_{j_2}^{(j_3)})B_{kj_2}B_{j_3j_1} + \right.$$

$$\left.\frac{1}{2}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}(\overline{\Psi}_{j_1} - \mu_{j_1})(\overline{\Psi}_{j_2} - \mu_{j_2})\sum_{j_3=1}^{p}B_{kj_3}\sum_{j_4=1}^{p}\sum_{j_5=1}^{p}B_{j_4j_2}\mu_{j_3}^{(j_4j_5)}B_{j_5j_1}\right\}_{k=1,\ldots,p}.$$

We now present the three proposals for $t_\mu$.

Proposal 1: $t_\mu = t_0 + E_h[e_T]$.

This requires us to compute a different adjustment $E_h[e_T]$ at each point the density is to be approximated. Note that the adjustment is a function of $t_0$ and its computation must be integrated into the procedure of solving the non-linear system for $\alpha$ and $t_0$. In addition, it follows from the discussion in Section 3.5 that

$$\begin{aligned} & E_h[e_{Tk}] \\ = \; & \frac{1}{n^2}\sum_{l=1}^{n}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{j_3=1}^{p}\{E_h[\Psi_{j_1 l}\Psi_{j_2 l}^{(j_3)}] - \mu_{j_1 l}\mu_{j_2 l}^{(j_3)}\}B_{kj_2}B_{j_3 j_1} + \\ & \frac{1}{2n^2}\sum_{l=1}^{n}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\{E_h[\Psi_{j_1 l}\Psi_{j_2 l}] - \mu_{j_1 l}\mu_{j_2 l}\}\sum_{j_3=1}^{p}B_{kj_3}\sum_{j_4=1}^{p}\sum_{j_5=1}^{p}B_{j_4 j_2}\mu_{j_3}^{(j_4 j_5)}B_{j_5 j_1}, \end{aligned}$$

$k = 1, \ldots, p$. Therefore to implement the adjustment, we need to evaluate

$$B_{j_1 j_2},\;\; \mu_{jl},\;\; \mu_{j_1 l}^{(j_2)},\;\; \mu_{j_1}^{(j_2 j_3)},\;\; E_h[\Psi_{j_1 l}\Psi_{j_2 l}],\;\; E_h[\Psi_{j_1 l}\Psi_{j_2 l}^{(j_3)}].$$

Note that the first three quantities are needed no matter if an adjustment is implemented or not, so that the additional computation for the adjustment is to evaluate the last three quantities.

Proposal 2: $t_\mu = t_0 + E_f[e_T]$.

This simplifies Proposal 1 by taking a constant adjustment over $t_0$. The adjustment is computed only once under $f$ and essentially does not increase the computational effort of the approximation. Note that $f$ is simply the special case of $h$ with $\alpha = 0$, the computation of the adjustment is basically the same as that in the first proposal.

The third adjustment is proposed particularly for an estimator that is defined via the Huber's score function $\Psi_c$. We first give a brief review of the score function and

define some notation for further discussions.

Define $\Psi_c(r) = max\{-c, min\{c, r\}\}$, where $c$ is a real constant. Then the first two derivatives of $\Psi_c(r)$ are

$$\Psi_c'(r) = I_c(r), \quad \Psi_c''(r) = \delta_{-c}(r) - \delta_c(r) \equiv \delta_\Psi(r),$$

where $I_c(r)$ equals 1 if $|r| < c$, and 0 otherwise, $\delta_c(r)$ is the Dirac delta function and may be defined by the relation

$$\int_r \delta_c(r)u(r)dr = u(c) \tag{A.1}$$

for any continuous function $u$ (see Kukin, 1989, page 41).

Define

$$\Psi_{trc}(r) = \Psi_c(r)I_c(r) = rI_c(r),$$

$$I_{trc}(r) = I_c(r)I_c(r) = I_c(r), \quad \delta_{trc}(r) = \delta_\Psi(r)I_c(r) = 0,$$

and in general, for any $u = u(\Psi_c, I_c, \delta_\Psi)$,

$$u_{trc} = u(\Psi_{trc}, I_{trc}, \delta_{trc}).$$

The third adjustment is as follows.

Proposal 3: $t_\mu = t_0 + E_h[e_{Ttrc}]$.

This is a refinement of the first proposal such that for $1 \leq k \leq p$,

$$\begin{aligned} & E_h[e_{Tk\,trc}] \\ = \; & E_h\left[\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{j_3=1}^{p}(\overline{\Psi}_{j_1trc} - \mu_{j_1trc})(\overline{\Psi}_{j_2trc}^{(j_3)} - \mu_{j_2trc}^{(j_3)})B_{kj_2}B_{j_3j_1} + \right. \\ & \left. \frac{1}{2}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}(\overline{\Psi}_{j_1trc} - \mu_{j_1trc})(\overline{\Psi}_{j_2trc} - \mu_{j_2trc})\sum_{j_3=1}^{p}B_{kj_3}\sum_{j_4=1}^{p}\sum_{j_5=1}^{p}B_{j_4j_2}\mu_{j_3trc}^{(j_4j_5)}B_{j_5j_1}\right] \end{aligned}$$

$$= \frac{1}{n^2}\sum_{l=1}^{n}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{j_3=1}^{p}\{E_h[\Psi_{j_1 l trc}\Psi_{j_2 l trc}^{(j_3)}] - \mu_{j_1 l trc}\mu_{j_2 l trc}^{(j_3)}\}B_{kj_2}B_{j_3 j_1} +$$

$$\frac{1}{2n^2}\sum_{l=1}^{n}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\{E_h[\Psi_{j_1 l trc}\Psi_{j_2 l trc}] - \mu_{j_1 l trc}\mu_{j_2 l trc}\} \times$$

$$\sum_{j_3=1}^{p} B_{kj_3}\sum_{j_4=1}^{p}\sum_{j_5=1}^{p} B_{j_4 j_2}\mu_{j_3 trc}^{(j_4 j_5)}B_{j_5 j_1}.$$

To implement the adjustment, we need to evaluate

$$B_{j_1 j_2},\ \mu_{j l trc},\ \mu_{j_1 l trc}^{(j_2)},\ \mu_{j_1 trc}^{(j_2 j_3)},\ E_h[\Psi_{j_1 l trc}\Psi_{j_2 l trc}],\ E_h[\Psi_{j_1 l trc}\Psi_{j_2 l trc}^{(j_3)}].$$

The computational requirement for these quantities are similar to that for the first proposal. We postpone the discussion to the next section.

## A.2 Adjustments for Huber-type estimators

We now derive the adjustments proposed in Section A.1 for the Huber-type estimators. The derivation is based on the regression model of Section 2.2.

Recall that the $p$-dimensional score functions for the estimators are

$$\Psi_l(r_l) = \begin{bmatrix} \Psi_c(r_l)X_{l1} \\ \vdots \\ \Psi_c(r_l)X_{lp-1} \\ \Psi_c^2(r_l) - \beta \end{bmatrix} = \begin{bmatrix} \Psi_c(r_l)X_{lj} \\ \Psi_c^2(r_l) - \beta \end{bmatrix}_{1\le j\le p-1},$$

where

$$r_l = \frac{Y_l - X_l^T\theta}{\sigma},$$

$l = 1,\ldots,n$. Note that the last matrix representation is adopted in the discussion in this section. To evaluate the adjustments, we also need the following matrices.

1. The first order partial derivatives of the score functions.

$$\frac{\partial\Psi_l(r_l)}{\partial\eta^T} = -\frac{1}{\sigma}\begin{bmatrix} \Psi_c'(r_l)X_{lj_1}X_{lj_2} & \Psi_c'(r_l)r_lX_{lj_1} \\ 2\Psi_c'(r_l)\Psi_c(r_l)X_{lj_2} & 2\Psi_c'(r_l)\Psi_c(r_l)r_l \end{bmatrix}_{1\le j_1,j_2\le p-1}$$

$$= -\frac{1}{\sigma}\begin{bmatrix} I_c(r_l)X_{lj_1}X_{lj_2} & \Psi_{trc}(r_l)X_{lj_1} \\ 2\Psi_{trc}(r_l)X_{lj_2} & 2\Psi_{trc}^2(r_l) \end{bmatrix}_{1\le j_1,j_2\le p-1}.$$

2. The second order partial derivatives of the score functions.

$$\frac{\partial^2\Psi_{j_1l}(r_l)}{\partial\eta\partial\eta^T}$$

$$= \frac{1}{\sigma^2}\begin{bmatrix} \Psi_c''(r_l)X_{lj_1}X_{lj_2}X_{lj_3} & \{\Psi_c''(r_l)r_l + \Psi_c'(r_l)\}X_{lj_1}X_{lj_2} \\ \{\Psi_c''(r_l)r_l + \Psi_c'(r_l)\}X_{lj_1}X_{lj_3} & \{\Psi_c''(r_l)r_l^2 + 2\Psi_c'(r_l)r_l\}X_{lj_1} \end{bmatrix}_{1\le j_2,j_3\le p-1}$$

$$= \frac{1}{\sigma^2}\begin{bmatrix} \delta_\Psi(r_l)X_{lj_1}X_{lj_2}X_{lj_3} & \{\delta_\Psi(r_l)r_l + I_c(r_l)\}X_{lj_1}X_{lj_2} \\ \{\delta_\Psi(r_l)r_l + I_c(r_l)\}X_{lj_1}X_{lj_3} & \{\delta_\Psi(r_l)r_l^2 + 2\Psi_{trc}(r_l)\}X_{lj_1} \end{bmatrix}_{1\le j_2,j_3\le p-1},$$

$j_1 = 1, \ldots, p-1$, and

$$\frac{\partial^2 \Psi_{pl}(r_l)}{\partial \eta \partial \eta^T} = \frac{2}{\sigma^2} \left[ \begin{array}{c} \{\Psi_c''(r_l)\Psi_c(r_l) + \Psi_c'^2\} X_{lj_2} X_{lj_3} \\ \{\Psi_c''(r_l)\Psi_c(r_l) r_l + \Psi_c'^2(r_l) r_l + \Psi_c'(r_l)\Psi_c(r_l)\} X_{lj_3} \end{array} \right.$$

$$\left. \begin{array}{c} \{\Psi_c''(r_l)\Psi_c(r_l) r_l + \Psi_c'^2(r_l) r_l + \Psi_c'(r_l)\Psi_c(r_l)\} X_{lj_2} \\ \Psi_c''(r_l)\Psi_c(r_l) r_l^2 + \Psi_c'^2(r_l) r_l^2 + 2\Psi_c'(r_l)\Psi_c(r_l) r_l \end{array} \right]_{1 \le j_2, j_3 \le p-1}$$

$$= \frac{2}{\sigma^2} \left[ \begin{array}{c} \{\delta_\Psi(r_l)\Psi_c(r_l) + I_c(r_l)\} X_{lj_2} X_{lj_3} \\ \{\delta_\Psi(r_l)\Psi_c(r_l) r_l + 2\Psi_{trc}(r_l)\} X_{lj_3} \end{array} \right.$$

$$\left. \begin{array}{c} \{\delta_\Psi(r_l)\Psi_c(r_l) r_l + 2\Psi_{trc}(r_l)\} X_{lj_2} \\ \delta_\Psi(r_l)\Psi_c(r_l) r_l^2 + 3\Psi_{trc}^2(r_l) \end{array} \right]_{1 \le j_2, j_3 \le p-1}.$$

3. The product of the score functions.

$$\Psi_l(r_l)\Psi_l^T(r_l) = \left[ \begin{array}{cc} \Psi_c^2(r_l) X_{lj_1} X_{lj_2} & \Psi_c(r_l)\{\Psi_c^2(r_l) - \beta\} X_{lj_1} \\ \Psi_c(r_l)\{\Psi_c^2(r_l) - \beta\} X_{lj_2} & \{\Psi_c^2(r_l) - \beta\}^2 \end{array} \right]_{1 \le j_1, j_2 \le p-1}.$$

4. The product of the score functions and the first order partial derivatives.

$$\Psi_{j_1 l}(r_l) \frac{\partial \Psi_l(r_l)}{\partial \eta} = -\frac{\Psi_c(r_l) X_{lj_1}}{\sigma} \left[ \begin{array}{cc} I_c(r_l) X_{lj_2} X_{lj_3} & \Psi_{trc}(r_l) X_{lj_2} \\ 2\Psi_{trc}(r_l) X_{lj_3} & 2\Psi_{trc}^2(r_l) \end{array} \right]_{1 \le j_2, j_3 \le p-1}$$

$$= -\frac{1}{\sigma} \left[ \begin{array}{cc} \Psi_{trc}(r_l) X_{lj_1} X_{lj_2} X_{lj_3} & \Psi_{trc}^2(r_l) X_{lj_1} X_{lj_2} \\ 2\Psi_{trc}^2(r_l) X_{lj_1} X_{lj_3} & 2\Psi_{trc}^3(r_l) X_{lj_1} \end{array} \right]_{1 \le j_2, j_3 \le p-1},$$

$j_1 = 1, \ldots, p-1$, and

$$\Psi_{pl}(r_l) \frac{\partial \Psi_l(r_l)}{\partial \eta} = -\frac{\Psi_c^2(r_l) - \beta}{\sigma} \left[ \begin{array}{cc} I_c(r_l) X_{lj_2} X_{lj_3} & \Psi_{trc}(r_l) X_{lj_2} \\ 2\Psi_{trc}(r_l) X_{lj_3} & 2\Psi_{trc}^2(r_l) \end{array} \right]_{1 \le j_2, j_3 \le p-1}$$

$$= -\frac{1}{\sigma} \left[ \begin{array}{cc} \Psi_{trc}^2(r_l) X_{lj_2} X_{lj_3} & \Psi_{trc}^3(r_l) X_{lj_2} \\ 2\Psi_{trc}^3(r_l) X_{lj_3} & 2\Psi_{trc}^4(r_l) \end{array} \right]_{1 \le j_2, j_3 \le p-1} + \frac{\beta}{\sigma} \frac{\partial \Psi_l(r_l)}{\partial \eta}.$$

Taking expectation on these matrices at $\eta = t_0$ gives us the quantities to compute the adjustments of Proposals 1 and 2. For the third proposal, we replace $(\Psi_c, I_c, \delta_\Psi)$ by $(\Psi_{trc}, I_c, 0)$ in the matrices to obtain the required quantities.

We conclude this section with the following comments.

1. To compute the expectations of the matrices, we basically need to evaluate the expected values of

$$\Psi_c^k(r_l), \quad k = 1, \ldots, 4,$$

and the probabilities

$$P(-c < r_l < c), \quad P(r_l < -c) \quad \text{and} \quad P(r_l > c)$$

under $h$.

2. The evaluation of the seven quantities in the first comment can be done numerically. For a better efficiency and in most situations, the quantities can be simplified algebrically before the implementation for an application.

3. The computational requirement of Proposal 3 is slightly less than that of Proposal 1. In particular, we have

$$\mu_{l\,trc}^{(j)} = \mu_l^{(j)} \quad \text{and} \quad E_h[\Psi_{j_1 l\,trc}\Psi_{j_2 l\,trc}^{(j_3)}] = E_h[\Psi_{j_1 l}\Psi_{j_2 l}^{(j_3)}].$$

For the other expectations, note that in general we have

$$E_h[\Psi_c^k] = E_h[\Psi_{trc}^k] + (-c)^k P(r_l < -c) + c^k P(r_l > c).$$

4. For those functions involving $\delta_\Psi$, the expectations follow from the definition (A.1) of $\delta_c$. For example,

$$E_h\left[\delta_\Psi\left(\frac{Y_l - X_l^T t_\theta}{t_\sigma}\right)\right] = \int_{y_l}\left\{\delta_{-c}\left(\frac{y_l - X_l^T t_\theta}{t_\sigma}\right) - \delta_c\left(\frac{y_l - X_l^T t_\theta}{t_\sigma}\right)\right\} h_l(y_l)\, dy_l$$

$$= t_\sigma \int_{r_l} \{\hat{o}_{-c}(r_l) - \delta_c(r_l)\}\ h_l(r_l t_\sigma + X_l^T t_\theta)\ dr_l$$

$$= t_\sigma \{h_l(-c\,t_\sigma + X_l^T t_\theta) - h_l(c\,t_\sigma + X_l^T t_\theta)\},$$

and generally

$$E_h\left[\delta_\Psi\left(\frac{Y_l - X_l^T t_\theta}{t_\sigma}\right) u\left(\frac{Y_l - X_l^T t_\theta}{t_\sigma}\right)\right]$$

$$= t_\sigma \int_{r_l} \{\delta_{-c}(r_l) - \delta_c(r_l)\}\ u(r_l)\ h_l(r_l t_\sigma + X_l^T t_\theta)\ dr_l$$

$$= t_\sigma \{u(-c)\ h_l(-c\,t_\sigma + X_l^T t_\theta) - u(c)\ h_l(c\,t_\sigma + X_l^T t_\theta)\}.$$

# Appendix B

# Ronchetti's $\tau$-test

## B.1 Definition and asymptotic distribution

For simplicity, consider the linear model of Section 2.2 with $\sigma = 1$ and $\theta$ being $p$-dimensional. Suppose that we want to test the hypothesis

$$H : \theta_{q+1} = \cdots = \theta_p = 0, \;\; 0 < q < p,$$

Ronchetti (see Hampel et al., 1986, page 346) proposes the following class of tests that can be viewed as an extension of the log-likelihood ratio test for linear models.

**Definition B.1** *Define the corresponding $M$-estimators $\hat{\theta}_F$ and $\hat{\theta}_R$ in the full and reduced model, respectively, by*

$$\Gamma(\hat{\theta}_F) = \min\{\Gamma(\theta)|\theta \in \Theta\}, \quad \Gamma(\hat{\theta}_R) = \min\{\Gamma(\theta)|\theta \in \Theta_R\},$$

*where $\Theta_R$ is the subspace of the parameter space $\Theta$ obtained by imposing the condition $H$ and*

$$\Gamma(\theta) = \sum_{i=1}^{n} \tau\left\{X_i, Y_i - X_i^T\theta\right\}.$$

*A $\tau$-test is a test based on the test statistic*

$$T_\tau = \frac{2}{p-q}\frac{1}{n}\{\Gamma(\hat{\theta}_R) - \Gamma(\hat{\theta}_F)\}.$$

□

In addition, Ronchetti (see Hampel et al., 1986, page 352) derives the asymptotic distribution of $nT_\tau$ when $X_l$'s are independent and identically distributed. For instance, the asymptotic distribution under the hypothesis $H$ is the distribution of

$$\frac{1}{p-q}\sum_{j=q+1}^{p}\lambda_j\chi_j^2,$$

where $\chi_j^2$'s are independent standard $\chi^2$ random variables with one degree of freedom, and $\lambda_j$'s are the $p-q$ positive eigenvalues of $Q(M^{-1} - M_R^-)$,

$$Q = E_f\left[\left\{\left.\frac{\partial\tau(X_1,\varepsilon_1)}{\partial\varepsilon_1}\right|_{\varepsilon_1=Y_1-X_1^T\theta_0}\right\}^2 X_1X_1^T\right],$$

$$M = E_f\left[\left.\frac{\partial^2\tau(X_1,\varepsilon_1)}{\partial\varepsilon_1^2}\right|_{\varepsilon_1=Y_1-X_1^T\theta_0} X_1X_1^T\right], \quad M_R^- = \begin{bmatrix} M_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix},$$

$M_{11}$ being the upper-left corner of $M$ of order $q$ by $q$. To compute the tail probabilities of the asymptotic distribution, we can apply the approximation for linear combinations of $\chi^2$ random variables by Field (1993).

## B.2 An unsolved problem

To illustrate the problem that we encounter in applying our density approximation to the $\tau$-test, we consider the Huber-type estimator for the model in Section B.1. We now have

$$\tau(X_l, r_l(\theta)) = \frac{1}{2}\Psi_c(r_l(\theta))\{2r_l(\theta) - \Psi_c(r_l(\theta))\}, \quad l = 1, \ldots, n,$$

where $r_l(\theta) = Y_l - X_l^T\theta$ and $\Psi_c$ is the Huber's score function, so that

$$\frac{\partial\tau(X_l, r_l(\theta))}{\partial\theta} = -\Psi_c(r_l(\theta))X_l.$$

Let $X_R$ be the first $q$ columns of $X$. The $M$-estimators are defined as the solutions of

$$\frac{1}{n}\sum_{l=1}^{n}\Psi_c(r_l(\hat{\theta}_F))X_{lj} = 0, \quad j = 1, \ldots, p, \tag{B.1}$$

$$\frac{1}{n}\sum_{l=1}^{n}\Psi_c(r_l(\hat{\theta}_R))X_{Rlj} = 0, \quad j = 1, \ldots, q. \tag{B.2}$$

By convention, we define $r_l(\hat{\theta}_R) = Y_l - {X_{Rl}}^T\hat{\theta}_R$.

Now, consider the approximation of the densities of $T_\tau$. Note that $T_\tau$ is a function involving the random observations, we cannot directly apply the density approximation. To tackle the problem, we define $T_\tau$ as the solution of

$$\frac{1}{n}\sum_{l=1}^{n}\left\{\frac{2}{p-q}\{\tau(X_l, r_l(\hat{\theta}_R)) - \tau(X_l, r_l(\hat{\theta}_F))\} - T_\tau\right\} = 0 \tag{B.3}$$

and consider $\hat{\eta} = (\hat{\theta}_F, \hat{\theta}_R, T_\tau)$ as the simultaneous solution of the $(p+q+1)$-system consisting of (B.1), (B.2) and (B.3). To approximate $\hat{\eta}$, we have

$$A = -\frac{1}{n}\sum_{l=1}^{n}E_h\left[\begin{array}{ccc|c} I_c(r_l(\hat{\theta}_F))X_lX_l^T & 0 & 0 & \\ 0 & I_c(r_l(\hat{\theta}_R))X_{Rl}{X_{Rl}}^T & 0 & \\ -2(p-q)^{-1}\Psi_c(r_l(\hat{\theta}_F))X_l^T & 2(p-q)^{-1}\Psi_c(r_l(\hat{\theta}_R))X_{Rl}{}^T & 1 & \end{array}\right|_{\hat{\eta}=t_0}$$

$$= -\frac{1}{n}\sum_{l=1}^{n}\begin{bmatrix} E_h[I_c(r_l(t_F))]X_l X_l^T & 0 & 0 \\ 0 & E_h[I_c(r_l(t_R))]X_{Rl}X_{Rl}{}^T & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The last equality follows from the conditions for choosing $t_0 = (t_F, t_R, t_\tau)$. Therefore

$$B = \begin{bmatrix} n\left\{\sum_{l=1}^{n} E_h[I_c(r_l(t_F))]X_l X_l^T\right\}^{-1} & 0 & 0 \\ 0 & n\left\{\sum_{l=1}^{n} E_h[I_c(r_l(t_R))]X_{Rl}X_{Rl}{}^T\right\}^{-1} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and the linear approximation

$$\begin{aligned} G &= t_\tau + \frac{1}{n}\sum_{l=1}^{n}\left\{\frac{2}{p-q}\{\tau(X_l, r_l(t_R)) - \tau(X_l, r_l(t_F))\} - t_\tau\right\} \\ &= \frac{2}{p-q}\frac{1}{n}\sum_{l=1}^{n}\{\tau(X_l, r_l(t_R)) - \tau(X_l, r_l(t_F))\} \end{aligned}$$

which is simply the first term of a Taylor series expansion for $T_\tau$.

Our problem is that when $G$ is evaluated under $f$, we have $\theta_0 = (\theta_{R0}, 0)$, $t_F = \theta_0$, $t_R = \theta_{R0}$ and therefore

$$G = \frac{2}{p-q}\frac{1}{n}\sum_{l=1}^{n}\{\tau(X_l, r_l(\theta_{R0})) - \tau(X_l, r_l(\theta_0))\}$$

which vanishes. This violates our assumption A8 in Section 3.2 and we cannot proceed to obtain an approximation. In fact, by expanding $\tau(X_l, r_l(\hat{\theta}_R)) - \tau(X_l, r_l(\hat{\theta}_F))$ about $\theta_0$, we can easily see that the difference is determined by the second term and up and the first term approximation is always zero.

## B.3 A potential solution

In a recent conversation, Ronchetti suggested that we approximate the densities of a quadratic term in an expansion of $T_\tau$ rather than the densities of the test statistic itself. Precisely, he shows in Hampel et al. (1986, page 352) that

$$(p-q)nT_\tau = V^T(\theta_0)\,(M^{-1} - M_R^-)\,V(\theta_0) + \cdots,$$

where

$$V(\theta_0) = \frac{1}{\sqrt{n}} \sum_{l=1}^{n} \Psi_c(r_l(\theta_0))X_l.$$

Note that the quadratic term is the basis used by Ronchetti (see Hampel et al., Chapter 7) to derive the asymptotic distribution of $T_\tau$.

To approximate the densities of the quadratic term, we suppose that the matrix $M^{-1} - M_R^-$ of rank $p-q$ has non-zero eigenvalues $\lambda_i$ and corresponding eigenvectors $a_i$, $i = 1, \ldots, p-q$. Then

$$\begin{aligned} M^{-1} - M_R^- &= \left[\sqrt{\lambda_1}a_1 \;\; \cdots \;\; \sqrt{\lambda_{p-q}}a_{p-q}\right] \begin{bmatrix} \sqrt{\lambda_1}a_1^T \\ \vdots \\ \sqrt{\lambda_{p-q}}a_{p-q}^T \end{bmatrix} \\ &\equiv LL^T. \end{aligned}$$

Writing

$$\begin{aligned} \sqrt{n}L^TV(\theta_0) &= \begin{bmatrix} \sum_{l=1}^n \Psi_c(r_l(\theta_0))\,(\sum_{i=1}^p X_{li}L_{i1}) \\ \vdots \\ \sum_{l=1}^n \Psi_c(r_l(\theta_0))\,(\sum_{i=1}^p X_{li}L_{ip-q}) \end{bmatrix} \\ &\equiv U, \end{aligned}$$

we can transform the quadratic term to a sum of squares, that is,

$$nV^T(\theta_0)\,(M^{-1} - M_R^-)\,V(\theta_0) = nV^T(\theta_0)LL^TV(\theta_0) = U^TU = \sum_{i=1}^{p-q} U_i^2.$$

Since $U_t$'s are just linear combinations of the score functions, we can apply the result in Chapter 5 and expect to obtain accurate approximations of their joint densities.

To compute the distribution of the $\tau$-test from the joint density approximation, we require a $(p-q)$-dimensional numerical integration. This may not be very attractive if $p-q$ is large. An alternative could be to generate the joint densities of $U_t^2$'s from that of $U_t$'s and then apply the result of DiCiccio and Martin (1991). Note that a direct application of their result to the joint densities of $U_t$'s has some problems since the gradient of the sum of squares vanishes at the maximum (see DiCiccio and Martin, 1991). The possibility is now under research.

A final comment to this potential solution is that we approximate the densities of a quadratic approximation for $T_\tau$. The performance is not clear at the moment. However, since the asymptotic distribution is derived using the quadratic term, we expect our approximation at least to improve over the asymptotic result.

# Appendix C

# Sample programs

## C.1 Marginal density approximation using univariate $G_p$

```
c This program generates the numerical results in Tables 4.1a & 4.1b

c Model     : Location-scale
c Estimator : Huber-type
c rho(eta)  : theta

c Main program
c       program margdens.for
           implicit double precision (a-h,o-z)
           common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1            n0,ir
           parameter (nitv=450,nsto=300,nftr=1,ditv=1.d-2,dsto=1.d-2)
           dimension px(2,nsto)
           call init_1
c Computing the density approximation over a grid of points
           do 100 ix = 1, nitv
              call init_2 ( ix, ditv )
              call compute_us
              call init_3
              call compute_rk
              call compute_px ( ix, px, nsto, nftr )
100        continue
```

```
         call store_px ( px, nsto, ditv, dsto )
         close ( 99 )
      end

c General initialization
      subroutine init_1
         implicit double precision (a-h,o-z)
         common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1          n0,ir
         n0    = 10
         c0    = 1.345
         b0    = 1 - 2 * ( c0 * dnorm ( c0 ) +
     1                     ( 1 - c0 ** 2 ) * s15acf ( c0, ifail ) )
         a0(1) = 0.d0
         a0(2) = 0.d0
         cp(1) = 0.d0
         cp(2) = 0.d0
         ta(2) = 1.d0
         return
      end

c Initialization before computing the approximation at each point
      subroutine init_2 ( ix, ditv )
         implicit double precision (a-h,o-z)
         common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1          n0,ir
         d0    = 1.d0
         ta(1) = ditv * ix
         return
      end

c Computing alpha and t_0
      subroutine compute_us
         implicit double precision (a-h,o-z)
         common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1          n0,ir
         parameter (xtol=1.d-10)
         dimension fp(2),tp(2),wa(19)
         external dpsi
         ifail = 0
         tp(1) = a0(1)
         tp(2) = ta(2)
```

```
          call c05nbf ( dpsi, 2, tp, fp, xtol, wa, 19, ifail )
          a0(1) = tp(1)
          ta(2) = tp(2)
          return
        end

c Quantities for computing the approximation
        subroutine init_3
          implicit double precision (a-h,o-z)
          common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1           n0,ir
          parameter (epsa=1.d-10,epsr=1.d-10,lw=4000,liw=1000)
          dimension w(lw),iw(liw)
          external dhx,a11,a12,a21,a22
          ifail = 0
          bd(1) = - c0 * ta(2) + ta(1)
          bd(2) = c0 * ta(2) + ta(1)
          c2b   = c0 ** 2 - b0
          phx1  = dexp ( - a0(1) * c0 + a0(2) * c2b ) * pfx ( bd(1) )
          phx2  = dexp ( a0(1) * c0 + a0(2) * c2b ) *
     1              ( 1 - pfx ( bd(2) ) )
          call d01ajf ( dhx, bd(1), bd(2), epsa, epsr, f0, abserr,
     1                  w, lw, iw, liw, ifail )
          d0    = phx1 + f0 + phx2
c Expectations of the first order partial derivatives
          call d01ajf ( a11, bd(1), bd(2), epsa, epsr, f1, abserr,
     1                  w, lw, iw, liw, ifail )
          call d01ajf ( a12, bd(1), bd(2), epsa, epsr, f2, abserr,
     1                  w, lw, iw, liw, ifail )
          call d01ajf ( a21, bd(1), bd(2)  epsa, epsr, f3, abserr,
     1                  w, lw, iw, liw, ifail )
          call d01ajf ( a22, bd(1), bd(2), epsa, epsr, f4, abserr,
     1                  w, lw, iw, liw, ifail )
          dt      = ta(2) / 2 / ( f1 * f4 - f2 * f3 )
c Matrix B
          bt(1,1) = 2 * f4 * dt
          bt(1,2) = - 2 * f3 * dt
          bt(2,1) = - f2 * dt
          bt(2,2) = f1 * dt
          gc(1)   = gx ( bd(1) )
          gc(2)   = gx ( bd(2) )
          return
```

```
      end

c Moments of Gp
      subroutine compute_rk
         implicit double precision (a-h,o-z)
         common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1          n0,ir
         parameter (epsa=1.d-10,epsr=1.d-10,lw=4000,liw=1000)
         dimension w(lw),iw(liw)
         external grhx
         ifail = 0
         c2b   = c0 ** 2 - b0
         phx1  = dexp ( - a0(1) * c0 + a0(2) * c2b ) *
     1           pfx ( bd(1) ) / d0
         phx2  = dexp ( a0(1) * c0 + a0(2) * c2b ) *
     1           ( 1 - pfx ( bd(2) ) ) / d0
         rk(1) = 0.d0
c        do 300 ix = 1,4
         do 300 ix = 1,2
            ir = ix
            call d01ajf ( grhx, bd(1), bd(2), epsa, epsr, fi, abserr,
     1                    w, lw, iw, liw, ifail )
            rk(ir) = ( gc(1) - rk(1) ) ** ir * phx1 + fi +
     1             ( gc(2) - rk(1) ) ** ir * phx2
300      continue
         rk(3) = rk(3) / rk(2) ** 1.5
         rk(4) = rk(4) / rk(2) ** 2 - 3
         return
      end

c Edgeworth density approximation
      subroutine compute_px ( ix, px, nsto, nftr )
         implicit double precision (a-h,o-z)
         common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1          n0,ir
         parameter (pi=3.141592653589793)
         dimension px(2,nsto)
         pr1   = dsqrt ( n0 / 2 / pi / rk(2) ) * d0 ** n0
         pr2   = pr1 * ( 1 + (rk(4) / 8 - 5 * rk(3) ** 2 / 72) / n0 )
         cp(1) = cp(1) + pr1
         cp(2) = cp(2) + pr2
         iy    = ix / nftr
```

```
            if ( ( iy .le. nsto ) .and. ( ( iy * nftr ) .eq. ix ) ) then
               px(1,iy) = cp(1)
               px(2,iy) = cp(2)
            endif
            return
          end

c Saving the numerical results
          subroutine store_px ( px, nsto, ditv, dsto )
            implicit double precision (a-h,o-z)
            common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1             n0,ir
            dimension px(2,nsto)
            cp(1) = cp(1) * 2
            cp(2) = cp(2) * 2
            open ( unit = 99, file = 'margdens.o21' )
            do 200 ix = 1, nsto
               px1 = 5.d-1 + px(1,ix) * ditv
c              px2 = 5.d-1 + px(2,ix) * ditv
               px3 = 5.d-1 + px(1,ix) / cp(1)
c              px4 = 5.d-1 + px(2,ix) / cp(2)
               write ( 99, * ) sngl ( dsto * ix ), sngl(px1), sngl(px3)
200         continue
            return
          end

c Quantities for the centering constraints
          subroutine dpsi ( np, tp, fp, iflag )
            implicit double precision (a-h,o-z)
            common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1             n0,ir
            parameter (epsa=1.d-10,epsr=1.d-10,lw=4000,liw=1000)
            dimension fp(np),tp(np),w(lw),iw(liw)
            external dpsi_1, dpsi_2
            ifail = 0
            a0(1) = tp(1)
            ta(2) = tp(2)
            bd(1) = - c0 * ta(2) + ta(1)
            bd(2) = c0 * ta(2) + ta(1)
            c2b   = c0 ** 2 - b0
c Tail probabilities under h
            phx1  = dexp ( - a0(1) * c0 + a0(2) * c2b ) *
```

```
     1            pfx ( bd(1) ) / d0
          phx2  = dexp ( a0(1) * c0 + a0(2) * c2b ) *
     1            ( 1 - pfx ( bd(2) ) ) / d0
          call d01ajf ( dpsi_1, bd(1), bd(2), epsa, epsr, f1, abserr,
     1                  w, lw, iw, liw, ifail )
          fp(1) = c0 * ( phx2 - phx1 ) + f1
          call d01ajf ( dpsi_2, bd(1), bd(2), epsa, epsr, f2, abserr,
     1                  w, lw, iw, liw, ifail )
          fp(2) = c2b * ( phx2 + phx1 ) + f2
          return
        end

c Integrand for the expectation of A_{11}
       function a11 ( x0 )
          implicit double precision (a-h,o-z)
          common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1           n0,ir
          a11 = dhx ( x0 )
          return
        end

c Integrand for the expectation of A_{12}
       function a12 ( x0 )
          implicit double precision (a-h,o-z)
          common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1           n0,ir
          a12 = dhx ( x0 ) * psi_1 ( x0 )
          return
        end

c Integrand for the expectation of A_{21}
       function a21 ( x0 )
          implicit double precision (a-h,o-z)
          common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1           n0,ir
          a21 = dhx ( x0 ) * psi_1 ( x0 )
          return
        end

c Integrand for the expectation of A_{22}
       function a22 ( x0 )
          implicit double precision (a-h,o-z)
```

```
      common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1       n0,ir
      a22 = dhx ( x0 ) * psi_1 ( x0 ) ** 2
      return
    end

c Conjugate density h
    function dhx ( x0 )
      implicit double precision (a-h,o-z)
      common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1       n0,ir
      parameter (pi=3.141592653589793)
c     dfx = dnorm ( x0 )
c     dfx = 0.9 * dnorm ( x0 ) + 0.01 * dnorm ( x0 / 1.d1 )
      dfx = 2 / dsqrt ( 3.d0 ) / pi / ( 1 + x0 ** 2 / 3 ) ** 2
c     dfx = 1 / pi / ( 1 + x0 ** 2 )
      dhx = dexp (a0(1) * psi_1 ( x0 ) + a0(2) * psi_2 ( x0 )) *
     1      dfx / d0
      return
    end

c Standard normal density
    function dnorm ( x0 )
      implicit double precision (a-h,o-z)
      common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1       n0,ir
      parameter (pi=3.141592653589793)
      dnorm = dexp ( - x0 ** 2 / 2 ) / dsqrt ( 2 * pi )
      return
    end

c Integrand for the expectation of Psi_1
    function dpsi_1 ( x0 )
      implicit double precision (a-h,o-z)
      common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1       n0,ir
      dpsi_1 = dhx ( x0 ) * psi_1 ( x0 )
      return
    end

c Integrand for the expectation of Psi_2
    function dpsi_2 ( x0 )
```

```
          implicit double precision (a-h,o-z)
          common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1           n0,ir
          dpsi_2 = dhx ( x0 ) * psi_2 ( x0 )
          return
        end

c Integrand for the moments of G_p
        function grhx ( x0 )
          implicit double precision (a-h,o-z)
          common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1           n0,ir
          grhx = ( gx ( x0 ) - rk(1) ) ** ir * dhx ( x0 )
          return
        end

c G_p
        function gx ( x0 )
          implicit double precision (a-h,o-z)
          common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1           n0,ir
          gx = psi_1 ( x0 ) * bt(1,1) + psi_2 ( x0 ) * bt(2,1)
          return
        end

c Error distribution
        function pfx ( x0 )
          implicit double precision (a-h,o-z)
          common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1           n0,ir
          parameter (pi=3.141592653589793)
          ifail = 0
c         pfx   = s15abf ( x0, ifail )
c         pfx   = 0.9 * s15abf ( x0, ifail ) + 0.1 *
c    1            s15abf ( x0/1.d1, ifail )
          pfx   = g01baf ( 3, x0, ifail )
c         pfx   = datan ( x0 ) / pi + 0.5
          return
        end

c First score function
        function psi_1 ( x0 )
```

```
          implicit double precision (a-h,o-z)
          common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1           n0,ir
          psi_1 = ( x0 - ta(1) ) / ta(2)
          return
        end

c Second score function
        function psi_2 ( x0 )
          implicit double precision (a-h,o-z)
          common b0,c0,d0,a0(2),bd(2),bt(2,2),cp(2),gc(2),rk(4),ta(2),
     1           n0,ir
          psi_2 = psi_1 ( x0 ) ** 2 - b0
          return
        end
```

## C.2 Tail probability approximation using multivariate $G_p$

```
c This program generates the numerical results in Tables 5.3

c Model     : Multiple regression
c Estimator : Huber-type or Mallow-type
c rho(eta)  : (theta_3, sigma) for studentized t-ratio

c tailprob.par - variable declaration for tailprob.for
          implicit double precision (a-h,o-z)
          parameter (c0=1.345,jp=4,js=jp+1,n0=20,ngrd=1000,
     1               mgrd=ngrd/2)
          common a0(2,js),axy(ngrd,ngrd,2,js),b0,bc(n0,2),bt(js,js),
     1           by(2),d0(n0),ditv(2),dmin,pc(n0,2),pxy(ngrd,ngrd),
     2           r1(2),ta(js),txy(ngrd,ngrd,js),w0(n0),x0(n0,jp),
     3           xa(n0),xcp(n0,4),yp(n0),ik,il,mby(2),nbx(2),
     4           nby(ngrd,2),id_ok

c Main program
c       program tailprob.for
          include 'tailprob.par'
          dimension dxy0(ngrd,ngrd)
          call init_gen ( igrd_ok, dxy0 )
          call init_grd ( igrd_ok, dxy0 )
          call comp_pxy
        end

c Computing alpha and t_0
        subroutine comp_at ( ig_ok )
          include 'tailprob.par'
          parameter (atmin=.1d-7,xtol=1.d-6,jf=js*3,
     1               nwa=jf*(3*jf+13)/2)
          dimension fp(jf),tp(jf),wa(nwa)
          external recenter
          ifail = 0
          ig_ok = 1
          if ( id_ok .eq. 1 ) then
            id_ix = r1(1) / ditv(1)
            id_iy = r1(2) / ditv(2)
            if ( id_ix * ditv(1) .lt. r1(1) ) id_ix = id_ix + 1
            if ( id_iy * ditv(2) .lt. r1(2) ) id_iy = id_iy + 1
            igx = id_ix + mgrd
```

```
            igy = id_iy + mgrd
            if ( id_iy .gt. nby(igx,2) ) then
               ig_ok = 0
            else
               do 120 i1 = 1, js
                  a0(1,i1) = axy(igx,igy,1,i1)
                  a0(2,i1) = axy(igx,igy,2,i1)
                  ta(i1)   = txy(igx,igy,i1)
120            continue
            endif
         endif
         if ( ig_ok .eq. 1 ) then
            do 100 i1 = 1, js
               tp(i1)      = a0(1,i1)
               tp(i1+js)   = a0(2,i1)
               tp(i1+js*2) = ta(i1)
100         continue
            call c05nbf (recenter, jf, tp, fp, xtol, wa, nwa, ifail)
            do 110 i1 = 1, js
               a0(1,i1) = tp(i1)
               a0(2,i1) = tp(i1+js)
               ta(i1)   = tp(i1+js*2)
               if ( dabs ( a0(1,i1) ) .lt. atmin ) a0(1,i1) = 0.d0
               if ( dabs ( a0(2,i1) ) .lt. atmin ) a0(2,i1) = 0.d0
               if ( dabs ( ta(i1) ) .lt. atmin ) ta(i1) = 0.d0
110            continue
         endif
         return
      end

c Computing the tail probabilities of the t-ratio
      subroutine comp_pxy
         include 'tailprob.par'
         parameter (aacc=1.d-6)
         dimension qtle(10),pval(9)
         external dxy,phi1,phi2

         qtle(1)  = 0.d0
c Quantiles of the t-ratio ( can be stored in a data file )
c Z
c        qtle(2)  = 5.17876d-1
c        qtle(3)  = 1.00345d0
c        qtle(4)  = 1.32808d0
c        qtle(5)  = 1.62123d0
c        qtle(6)  = 1.98737d0
```

```
c         qtle(7)  = 2.26441d0
c         qtle(8)  = 2.55690d0
c         qtle(9)  = 2.94193d0
c t3
          qtle(2)  = 5.32029d-1
          qtle(3)  = 1.02498d0
          qtle(4)  = 1.33818d0
          qtle(5)  = 1.63089d0
          qtle(6)  = 1.98264d0
          qtle(7)  = 2.25832d0
          qtle(8)  = 2.52483d0
          qtle(9)  = 2.90379d0
          qtle(10) = 1.d5
          ybnd1    = mby(1) * ditv(2)
          ybnd2    = mby(2) * ditv(2)
          do 200 iq = 1, 9
             by(1) = qtle(iq)
             by(2) = qtle(iq+1)
             call d01daf ( ybnd1, ybnd2, phi1, phi2, dxy, aacc,
     1                     pval(iq), npts, ifail )
             open ( unit = 98, file = 'tailprob.o21' )
             do 210 ip = 1, 9
210             write ( 98, * ) qtle(ip+1), pval(ip)
             close ( 98 )
200          continue
          return
       end

c General initialization
       subroutine init_gen ( igrd_ok, dxy0 )
          include 'tailprob.par'
          dimension dxy0(ngrd,ngrd)
          ifail = 0
          id_ok = 0
c Design matrix X and weight W
          open ( unit = 99, file = 'tailprob.par' )
          read ( 99, * ) ( ta(ip), ip = 1, jp )
          do 300 ix = 1, n0
             read ( 99, * ) ( x0(ix,ip), ip = 1, jp )
             w0(ix) = 1.d0
300       continue
          close ( 99 )
          nbx(1) = 0
          nbx(2) = 0
          do 310 ix = 1, ngrd
```

```
            nby(ix,1) = 0
            nby(ix,2) = 0
            do 320 iy = 1, ngrd
               pxy(ix,iy)  = 0.d0
               dxy0(ix,iy) = 0.d0
               do 330 ij = 1, js
                  axy(ix,iy,1,ij) = 0.d0
                  axy(ix,iy,2,ij) = 0.d0
                  txy(ix,iy,ij)   = 0.d0
330            continue
               txy(ix,iy,js) = 1.d0
320         continue
310      continue
         do 340 ij = 1, js
            a0(1,ij) = 0.d0
            a0(2,ij) = 0.d0
            ta(ij)   = 0.d0
340      continue
         ta(js)  = 1.d0
         b0      = ( 1 - 2 * ( c0 * dnorm ( c0 ) + ( 1 - c0 ** 2 ) *
     1             s15acf ( c0, ifail ) ) ) * ( n0 - jp ) / n0
         ditv(1) = 2.5d-1
         ditv(2) = 2.5d-1
         dmin    = 1.d-6
         it      = 3
         igrd_ok = 0
         open ( unit = 94, file = 'tailprob.t21' )
970      read ( 94, * ) ix, iy, tp_dxy
         if ( ix .lt. 0 ) goto 980
         if ( igrd_ok .eq. 0 ) igrd_ok = 1
         dxy0(ix,iy) = tp_dxy
         read ( 94, * ) ( axy(ix,iy,1,ij), ij = 1, js )
         read ( 94, * ) ( axy(ix,iy,2,ij), ij = 1, js )
         read ( 94, * ) ( txy(ix,iy,ij), ij = 1, js )
         goto 970
980      continue
         close ( 94 )
         return
      end

c Computing alpha and t_0 over a grid of points
      subroutine init_grd ( igrd_ok, dxy0 )
         include 'tailprob.par'
         dimension dxy0(ngrd,ngrd)
         if ( igrd_ok .eq. 0 )
```

```
     1         open ( unit = 94, file = 'tailprob.t21' )
           ity = -1
           do 400 i1 = 1, 2
              ixm     = 0
              iy      = 2 - i1
410           iy      = iy + ity
              igy     = iy + mgrd
              next_iy = 0
              ixm0    = ixm
              dxm     = 0.d0
              itx     = -1
              do 420 i2 = 1, 2
                 igx = ixm + mgrd
                 if ( i2 .eq. 2 ) then
                    do 430 ij = 1, js
                       a0(1,ij) = axy(igx,igy,1,ij)
                       a0(2,ij) = axy(igx,igy,2,ij)
                       ta(ij)   = txy(igx,igy,ij)
430                 continue
                 else if ( iy .ne. 0 ) then
                    do 440 ij = 1, js
                       a0(1,ij) = axy(igx,igy-ity,1,ij)
                       a0(2,ij) = axy(igx,igy-ity,2,ij)
                       ta(ij)   = txy(igx,igy-ity,ij)
440                 continue
                 endif
                 ix           = ixm + 2 - i2
450              ix           = ix + itx
                 igx          = ix + mgrd
                 nby(igx,i1) = iy
                 if ( igrd_ok .eq. 0 ) then
                    dxy0(igx,igy) = dxy ( ix * ditv(1), iy * ditv(2) )
                    do 460 ij = 1, js
                       axy(igx,igy,1,ij) = a0(1,ij)
                       axy(igx,igy,2,ij) = a0(2,ij)
                       txy(igx,igy,ij)   = ta(ij)
460                 continue
                    write ( 94, * ) igx, igy, dxy0(igx,igy)
                    write ( 94, * ) ( axy(igx,igy,1,ij), ij = 1, 2 )
                    write ( 94, * ) ( axy(igx,igy,1,ij), ij = 3, js )
                    write ( 94, * ) ( axy(igx,igy,2,ij), ij = 1, 2 )
                    write ( 94, * ) ( axy(igx,igy,2,ij), ij = 3, js )
                    write ( 94, * ) ( txy(igx,igy,ij), ij = 1, 2 )
                    write ( 94, * ) ( txy(igx,igy,ij), ij = 3, js )
                 endif
```

```
                if ( dxy0(igx,igy) .gt. dxm ) then
                   ixm0 = ix
                   dxm  = dxy0(igx,igy)
                endif
                if ( dxy0(igx,igy) .gt. dmin ) then
                   if ( next_iy .eq. 0 ) next_iy = 1
                   if ( itx .gt. 0 ) then
                      goto 450
                   else
                      itx = -itx
                   endif
                else
                   itx = -itx
                   if ( ( i2 .eq. 1 ) .and. ( nbx(1) .gt. ix ) ) then
                      nbx(1) = ix
                   else if ( ( i2 .eq. 2 ) .and. ( nbx(2) .lt. ix ) )
     1                 then
                      nbx(2) = ix
                   endif
                endif
                if ( iy * ditv(2) .lt. 1.d0 ) next_iy = 1
420          continue
             ixm = ixm0
             if ( ( next_iy .eq. 1 ) .and. ( ity .gt. 0 ) .and.
     1          ( igy .lt. ngrd ) ) goto 410
             ity     = -ity
             mby(i1) = iy
400       continue
          if ( igrd_ok .eq. 0 ) then
             write ( 94, * ) -1, -1, -1.d0
             close ( 94 )
          endif
          id_ok = 1
          return
       end

c Quantities under h_1 for computing the approximation
       subroutine para_1 ( xip0, xip1, xip2 )
          include 'tailprob.par'
          parameter (epsa=1.d-6,epsr=1.d-6,lw=4000,liw=1000)
          dimension w(lw),iw(liw)
          external dhx,dpsi_p1,dpsi_p2
          ifail  = 0
          xa(il) = 0.d0
          yp(il) = 0.d0
```

```
            do 500 i1 = 1, jp
               xa(il) = xa(il) + x0(il,i1) * ( a0(1,i1) + a0(2,i1) )
               yp(il) = yp(il) + x0(il,i1) * ta(i1)
500         continue
            bc(il,1) = -c0 * ta(js) + yp(il)
            bc(il,2) = c0 * ta(js) + yp(il)
            call d01ajf ( dhx, bc(il,1), bc(il,2), epsa, epsr, xip0,
     1                    aerr, w, lw, iw, liw, ifail )
            call d01ajf ( dpsi_p1, bc(il,1), bc(il,2), epsa, epsr, xip1,
     1                    aerr, w, lw, iw, liw, ifail )
            call d01ajf ( dpsi_p2, bc(il,1), bc(il,2), epsa, epsr, xip2,
     1                    aerr, w, lw, iw, liw, ifail )
            ac2b      = ( a0(1,js) + a0(2,js) ) * (c0**2 * w0(il) - b0)
            pc(il,1)  = dexp ( -xa(il) * w0(il) * c0 + ac2b ) *
     1                  pfx ( bc(il,1) )
            pc(il,2)  = dexp ( xa(il) * w0(il) * c0 + ac2b ) *
     1                  ( 1 - pfx ( bc(il,2) ) )
            d0(il)    = xip0 + pc(il,2) + pc(il,1)
            xcp(il,1) = ( xip1 + c0 * ( pc(il,2) - pc(il,1) ) ) / d0(il)
            xcp(il,2) = ( xip2 + c0 ** 2 * ( pc(il,2) + pc(il,1) ) ) /
     1                  d0(il)
            xip0      = xip0 / d0(il)
            xip1      = xip1 / d0(il)
            xip2      = xip2 / d0(il)
            return
          end

c Quantities for the centering constraints
          subroutine recenter ( jm, tp, fp, iflag )
            include 'tailprob.par'
            parameter (lwork=1000)
            dimension fp(jm),tp(jm),ipiv(js),work(lwork)
            do 600 i1 = 1, js
               a0(1,i1)    = tp(i1)
               a0(2,i1)    = tp(i1+js)
               ta(i1)      = tp(i1+js*2)
               fp(i1)      = 0.d0
               fp(i1+js)   = 0.d0
               fp(i1+js*2) = 0.d0
               do 610 i2 = 1, js
610               bt(i1,i2) = 0.d0
600         continue
            do 620 ix = 1, n0
               il = ix
               call para_l ( xip0, xip1, xip2 )
```

```
c Expectations of the first order partial derivatives and Psi
            do 630 i1 = 1, jp
               do 631 i2 = 1, jp
631               bt(i1,i2) = bt(i1,i2) + x0(il,i1) * x0(il,i2) *
     1                        w0(il) * xip0 / ta(js)
               bt(i1,js) = bt(i1,js) + x0(il,i1) * w0(il) * xip1 /
     1                     ta(js)
               fp(i1)    = fp(i1) + x0(il,i1) * w0(il) * xcp(il,1)
630         continue
            bt(js,js) = bt(js,js) + 2 * w0(il) * xip2 / ta(js)
            fp(js)    = fp(js) + xcp(il,2) * w0(il) - b0
620      continue
c Matrix B
         do 650 i1 = 1, jp
            do 660 i2 = 1, jp
660            bt(i1,i2) = bt(i1,i2) / n0
            bt(i1,js) = bt(i1,js) / n0
            bt(js,i1) = bt(i1,js) * 2
650      continue
         bt(js,js) = bt(js,js) / n0
         call f07adf ( js, js, bt, js, ipiv, info )
         call f07ajf ( js, bt, js, ipiv, work, lwork, info)
         do 670 i1 = 1, js
            if ( i1 .eq. ik ) then
               fp(ik+js) = ta(ik) - r1(1)
            else
               fp(i1+js) = a0(1,i1) * bt(ik,ik) - a0(1,ik) *
     1                     bt(ik,i1)
            endif
            if ( i1 .eq. js ) then
c Constant adjustment ( can be replaced by a subroutine )
c Z  -1.47302d-1
c t3 -1.11731d-1
               fp(js+js*2) = ta(js) - 1.11731d-1 - r1(2)
            else
               fp(i1+js*2) = a0(2,i1) * bt(js,js) - a0(2,js) *
     1                       bt(js,i1)
            endif
670      continue
         do 680 i1 = 1, jm
680         if ( dabs ( fp(i1) ) .lt. 1.d-15 ) fp(i1) = 0.d0
         return
       end

c Conjugate density h_l
```

```
      function dhx ( y0 )
         include 'tailprob.par'
         parameter (pi=3.141592653589793)
c        dfx = dnorm ( y0 )
         dfx = 2 / dsqrt ( 3.d0 ) / pi / ( 1 + y0 ** 2 / 3 ) ** 2
c        dfx = 0.9 * dnorm ( y0 ) + 0.01 * dnorm ( y0 / 1.d1 )
c        dfx = 1 / pi / ( 1 + y0 ** 2 )
         z0  = psi_c ( res_p ( y0 ) )
         dhx = dexp ( xa(il) * w0(il) * z0 + ( a0(1,js) +
     1         a0(2,js) ) * ( z0 ** 2 * w0(il) - b0 ) ) * dfx
         return
      end

c Max ( a, b )
      function dmaxi ( tp_1, tp_2 )
         include 'tailprob.par'
         if ( tp_1 .gt. tp_2 ) then
            dmaxi = tp_1
         else
            dmaxi = tp_2
         endif
         return
      end

c Min ( a, b )
      function dmini ( tp_1, tp_2 )
         include 'tailprob.par'
         if ( tp_1 .lt. tp_2 ) then
            dmini = tp_1
         else
            dmini = tp_2
         endif
         return
      end

c Standard normal density
      function dnorm ( z0 )
         include 'tailprob.par'
         parameter (pi=3.141592653589793)
         dnorm = dexp ( - z0 ** 2 / 2 ) / dsqrt ( 2 * pi )
         return
      end

c Integrand for the expectation of Psi_c
      function dpsi_p1 ( y0 )
```

```
         include 'tailprob.par'
         dpsi_p1 = dhx ( y0 ) * psi_c ( res_p ( y0 ) )
         return
      end

c Integrand for the expectation of Psi_c^2
      function dpsi_p2 ( y0 )
         include 'tailprob.par'
         dpsi_p2 = dhx ( y0 ) * psi_c ( res_p ( y0 ) ) ** 2
         return
      end

c Integrand for the expectation of Psi_c^3
      function dpsi_p3 ( y0 )
         include 'tailprob.par'
         dpsi_p3 = dhx ( y0 ) * psi_c ( res_p ( y0 ) ) ** 3
         return
      end

c Integrand for the expectation of Psi_c^4
      function dpsi_p4 ( y0 )
         include 'tailprob.par'
         dpsi_p4 = dhx ( y0 ) * psi_c ( res_p ( y0 ) ) ** 4
         return
      end

c Edgeworth density approximation
      function dxy ( x1, y1 )
         include 'tailprob.par'
         parameter (pi=3.141592653589793)
         parameter (epsa=1.d-6,epsr=1.d-6,lw=4000,liw=1000)
         dimension w(lw),iw(liw)
         external dpsi_p3,dpsi_p4
         ifail = 0
         r1(1) = x1
         r1(2) = dmaxi ( 5.d-2, y1 )
         call comp_at ( ig_ok )
         if ( ig_ok .eq. 0 ) then
            dxy = 0.d0
            goto 720
         endif
         var_1 = 0.d0
         var_2 = 0.d0
         cor12 = 0.d0
         p0    = 1.d0
```

```
      do 700 ix = 1, n0
         il  = ix
         p0  = p0 * d0(il)
         call d01ajf ( dpsi_p3, bc(il,1), bc(il,2), epsa, epsr,
     1                 xip3, aerr, w, lw, iw, liw, ifail )
         call d01ajf ( dpsi_p4, bc(il,1), bc(il,2), epsa, epsr,
     1                 xip4, aerr, w, lw, iw, liw, ifail )
         xcp(il,3) = ( xip3 + c0 ** 3 *
     1               ( pc(il,2) - pc(il,1) ) ) / d0(il)
         xcp(il,4) = ( xip4 + c0 ** 4 *
     1               ( pc(il,2) + pc(il,1) ) ) / d0(il)
         xbk = 0.d0
         xbs = 0.d0
         do 710 i1 = 1, jp
            xbk = xbk + x0(il,i1) * bt(ik,i1)
            xbs = xbs + x0(il,i1) * bt(js,i1)
710      continue
         xbk   = xbk * w0(il)
         xbs   = xbs * w0(il)
         bwk   = bt(ik,js) * w0(il)
         bws   = bt(js,js) * w0(il)
         x1_1  = xbk * xcp(il,1) + bwk * xcp(il,2)
         x1_2  = xbk ** 2 * xcp(il,2) + 2 * xbk * bwk *
     1           xcp(il,3) + bwk ** 2 * xcp(il,4)
         x2_1  = xbs * xcp(il,1) + bws * xcp(il,2)
         x2_2  = xbs ** 2 * xcp(il,2) + 2 * xbs * bws *
     1           xcp(il,3) + bws ** 2 * xcp(il,4)
         x12   = xbk * xbs * xcp(il,2) + ( xbk * bws +
     1           bwk * xbs ) * xcp(il,3) + bwk * bws * xcp(il,4)
         var_1 = var_1 + x1_2 - x1_1 ** 2
         var_2 = var_2 + x2_2 - x2_1 ** 2
         cor12 = cor12 + x12 - x1_1 * x2_1
700   continue
      dxy = p0 * n0 ** 2 / 2 / pi / dsqrt ( var_1 * var_2 -
     1      cor12 ** 2 )
720   continue
      return
    end

c Error distribution
      function pfx ( z0 )
         include 'tailprob.par'
         ifail = 0
c        pfx   = s15abf ( z0, ifail )
         pfx   = g01baf ( 3, z0, ifail )
```

```
c         pfx   = 0.9 * s15abf ( z0, ifail ) + 0.1 *
c    1              s15abf ( z0 / 1.d1, ifail )
c         pfx   = g01baf ( 1, z0, ifail )
          return
       end

c Lower limit for the double integral
       function phi1 ( pt_y )
          include 'tailprob.par'
          phi1 = dmini ( nbx(2) * ditv(1), by(1) * pt_y )
          return
       end

c Upper limit for the double integral
       function phi2 ( pt_y )
          include 'tailprob.par'
          phi2 = dmini ( nbx(2) * ditv(1), by(2) * pt_y )
          return
       end

c Huber's score function
       function psi_c ( z0 )
          include 'tailprob.par'
          psi_c = dmaxi ( -c0, dmini ( c0, z0 ) )
          return
       end

c Standardized value
       function res_p ( y0 )
          include 'tailprob.par'
          res_p = ( y0 - yp(il) ) / ta(js)
          return
       end
```

# Bibliography


[1] Bhattacharya, R.N. and Ghosh, J.K. (1978). On the validity of the formal Edgeworth expansion. *Annals of Statistics* **6**, 434-451.

[2] Daniels, H.E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics* **25**, 631-650.

[3] Daniels, H.E. and Young, G.A. (1991). Saddlepoint approximation for the studentized mean, with an application to the bootstrap. *Biometrika* **78**, 169-179.

[4] DiCiccio, T.J., Field, C.A., and Fraser, D.A.S. (1990). Approximations of marginal tail probabilities and inference for scalar parameters. *Biometrika* **77**, 77-95.

[5] DiCiccio, T.J. and Martin, M.A. (1991). Approximations of marginal tail probabilities for a class of smooth functions with applications to Bayesian and conditional inference. *Biometrika* **78**, 891-902.

[6] Edgeworth, F.Y. (1905). The law of error. *Proc. Cambridge Philos. Trans.* **20**, 36-66.

[7] Esseen, C.G. (1945). Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law. *Acta Mathematica* **77**, 1-125.

[8] Feller, W. (1971). *An introduction to probability theory and its applications* **2**, second edition. John Wiley & Sons, New York.

[9] Field, C.A. (1982). Small sample asymptotic expansions for multivariate *M*-estimates. *Annals of Statistics* **10**, 672-689.

[10] Field, C.A. (1993). Tail areas of linear combinations of chi-squares and non-central chi-squares. *Journal of Statistical Computation and Simulation* **45**, 243-248.

[11] Field, C.A. and Ronchetti, E.M. (1990). *Small sample asymptotics.* Institute of Mathematical Statistics, Hayward.

[12] Hall, P. (1992). *The Bootstrap and Edgeworth expansion.* Springer-Verlag, New York.

[13] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust statistics: The approach based on influence functions.* John Wiley & Sons, New York.

[14] Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and ANOVA. *American Statistician* **32**, 17-22.

[15] Hogg, R.V. and Craig, A.T. (1978). *Introduction to mathematical statistics*, fourth edition. Macmillan, New York.

[16] Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73-101.

[17] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics* **1**, 799-821.

[18] Huber, P.J. (1981). *Robust statistics.* John Wiley & Sons, New York.

[19] Kukin, V.D. (1989). Delta-functions. *Encyclopaedia of Mathematics* **3**. Kluwer Academic, Dordrecht.

[20] Kullback, S. (1959). *Information theory and statistics.* John Wiley & Sons, New York.

[21] Lawless, J.F. (1982). *Statistical models and methods for lifetime data.* John Wiley & Sons, New York.

[22] Lugannani, R. and Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability* **12**, 475-490.

[23] Maronna, R.A. and Yohai, V.J. (1981). Asymptotic behavior of general $M$-estimates for regression and scale with random carriers. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **58**, 7-20.

[24] McCullagh, P. (1987). *Tensor methods in statistics.* Chapman and Hall, London.

[25] Prakasa Rao, B.L.S. (1987). *Asymptotic theory of statistical inference.* John Wiley & Sons, New York.

[26] Staudte, R.G. and Sheather, S.J. (1990). *Robust estimation and testing.* John Wiley & Sons, New York.

[27] Tierney, L., Kass, R.E., and Kadane, J.B. (1989). Approximate marginal densities of nonlinear functions. *Biometrika* **76**, 425-433.

[28] Tingley, M.A. (1992). Small-sample intervals for regression. *Canadian Journal of Statistics* **20**, 271-280.

[29] Tingley, M.A. and Field, C.A. (1990). Small-sample confidence intervals. *Journal of the American Statistical Association* **85**, 427-434.

[30] Yohai, V.J. and Maronna, R.A. (1979). Asymptotic behavior of $M$-estimators for the linear model. *Annals of Statistics* **7**, 258-268.