Nonparametric Estimation of Distributions in a Large-p, Small-n Setting

Jeffrey D. Hart

Department of Statistics, Texas A&M University

Current and Future Trends in Nonparametrics

Columbia, South Carolina

October 11, 2007

Outline

- A random effects location model
- Multiple hypothesis testing
- Brief review of estimation results for location model
- Minimum distance estimation
- Simulation results
- Location-scale model
- Microarray example

Location model

$$X_{ij} = \mu_i + \sigma\epsilon_{ij}, \quad j = 1, \ldots, n,\ i = 1, \ldots, p.$$

Assumptions:

- $\mu_1, \ldots, \mu_p$ are i.i.d. as G.
- $\epsilon_{i1}, \ldots, \epsilon_{in}$, $i = 1, \ldots, p$, are i.i.d. as F, where F has mean 0 and standard deviation 1.
- All μ_i's independent of all ϵ_{ij}'s.
- σ is an unknown constant.

Problem of interest: *Obtain nonparametric estimates of F and G.*

Connection with deconvolution

When $n = 1$, our location model is a classic deconvolution model.

- In this case ($n = 1$), it's clear that F and G are not both identifiable.
- In deconvolution, the distribution of ϵ is assumed to be known, in which case it is possible to consistently estimate the distribution of μ from the X-data. [Carroll and Hall (1988, *JASA*)]

Multiple hypothesis testing

The **location model** is sometimes used in microarray analyses, where p is number of genes and n is number of measurements per gene.

- Test all hypotheses $H_{0i} : \mu_i = 0$, $i = 1, \ldots, p$.
- Typically, the distribution of a test statistic (under the null) will depend on F. Dependence on F is strong when n is small.
- Previous point implies that it is desirable to infer F.

Nonparametric estimation of F and G

- Is it possible to construct consistent nonparametric estimators of F and G when p goes to infinity but n is bounded?
- Perhaps surprisingly, the answer is "yes," even when n is as small as 2.

Two important early papers:

- Reiersøl (1950, *Econometrica*): Identifiability
- Wolfowitz (1957, *Ann. Math. Statist.*): Minimum distance estimation (MDE)

Two main types of estimators

- Explicit estimators – Based on characteristic function inversion, in analogy to simpler deconvolution problem.

- Minimum distance estimators – Choose F and G so that the induced distribution of $(X_{i1}, \ldots, X_{in})$ is a good match to the empirical distribution.

More recent literature

- Horowitz and Markatou (1996, *Review of Economic Studies*): Explicit estimators from panel data. (Error density assumed to be symmetric.)
- Li and Vuong (1998, *JMVA*): Explicit estimator in the location model.
- Hall and Yao (2003, *Ann. Statist.*): Explicit estimators and MDE histograms in location model.
- Neumann (2006): Strong consistency of MDEs of F_0 and G_0 in the location model.

Characteristic functions

$$\psi(s,t) = E\left[\exp\left(isX_{j1} + itX_{j2}\right)\right]$$

Under conditions more general than the location model, $\psi(s,t)$ is consistently estimated by

$$\hat{\psi}(s,t) = \binom{n}{2}^{-1} \sum_{1 \le j < k \le n} \hat{\psi}_{j,k}(s,t),$$

where

$$\hat{\psi}_{j,k}(s,t) = \frac{1}{p} \sum_{\ell=1}^{p} \exp\left(isX_{\ell j} + itX_{\ell k}\right).$$

In the location model,

$$\psi(s,t) = \psi_\mu(s+t)\psi_\epsilon(\sigma s)\psi_\epsilon(\sigma t).$$

An MDE metric

Suppose $\hat{\mu}_1, \ldots, \hat{\mu}_k$ are candidates for quantiles of G at probabilities $(j - 1/2)/k$, $j = 1, \ldots, k$.

An estimate of the cf of G is

$$\hat{\psi}_\mu(t) = \frac{1}{k} \sum_{j=1}^{k} e^{it\hat{\mu}_j}.$$

Given candidate quantiles for F, we may likewise compute an estimate $\hat{\psi}_\epsilon$ of ψ_ϵ.

Metric:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-h^2(s^2+t^2)} |\hat{\psi}(s,t) - \hat{\psi}_\mu(s+t)\hat{\psi}_\epsilon(\hat{\sigma}s)\hat{\psi}_\epsilon(\hat{\sigma}t)|^2 \, ds dt$$

An estimation algorithm

1. Compute metric for initial estimates of F_0 and G_0.
2. Randomly jitter initial quantiles of F_0, and recompute metric.
3. If new distance is smaller than the previous one, accept the jittered quantiles.
4. Repeat 2 and 3 some predetermined number of times.
5. Repeat 2-4 for estimates of the G_0 quantiles.
6. Iterate 2-5 until the distance changes by less than, say, 1% from one iteration to the next.

Simulations

$$X_{ij} = \mu_i + \sigma\epsilon_{ij}, \quad i = 1, \ldots, 1000, \; j = 1, 2$$

Two choices for G:

- Standard normal
- Bimodal mixture of two normals

Three choices for F:

- Standard normal
- Standard exponential shifted to have mean 0
- t_3-distribution rescaled to have variance 1

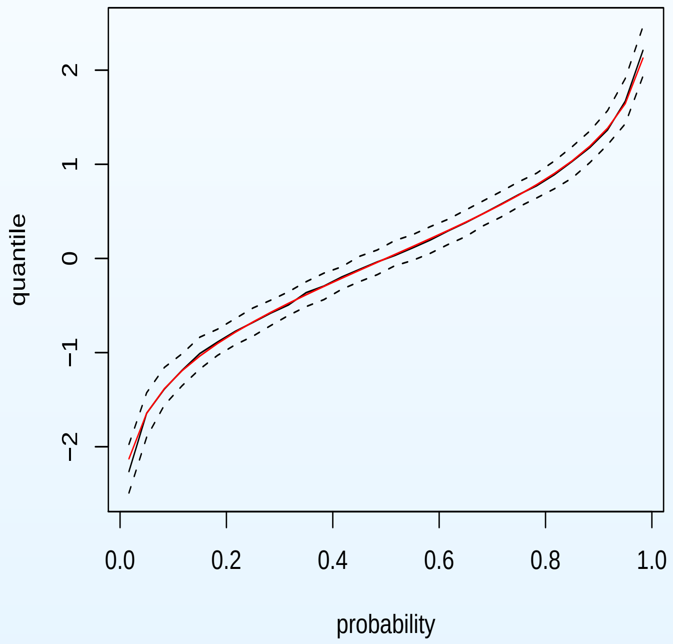
Two values of σ: 1 and 3

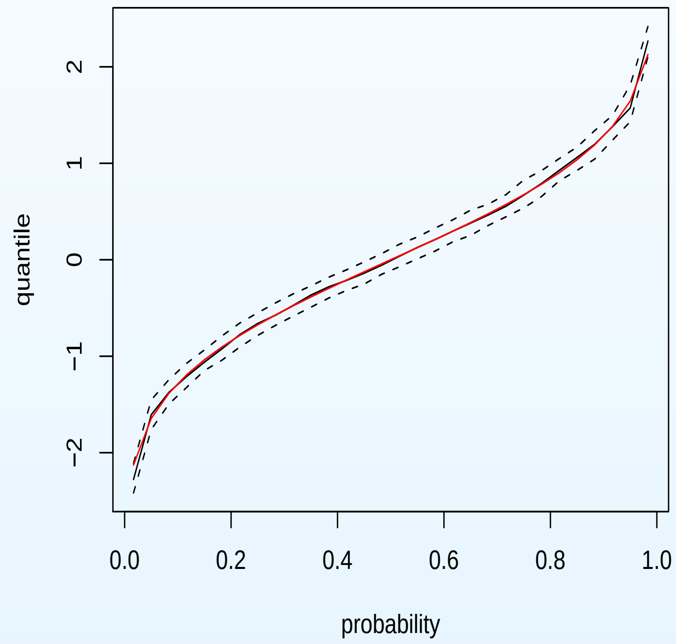
Simulations, continued

- Estimates of the quantile functions $G^{-1}(u)$ and $F^{-1}(u)$ were computed at $u = (j - 1/2)/30$, $j = 1, \ldots, 30$, for each data set generated from the location model.
- $\hat{\sigma}^2 = (2p)^{-1} \sum_{i=1}^{p} (X_{i1} - X_{i2})^2$
- Two hundred replications were performed at each combination of F, G and σ.
- Some of the results are summarized in the graphs to follow.

$G =$ Normal, $F =$ Normal

$\sigma = 1$ $\sigma = 3$



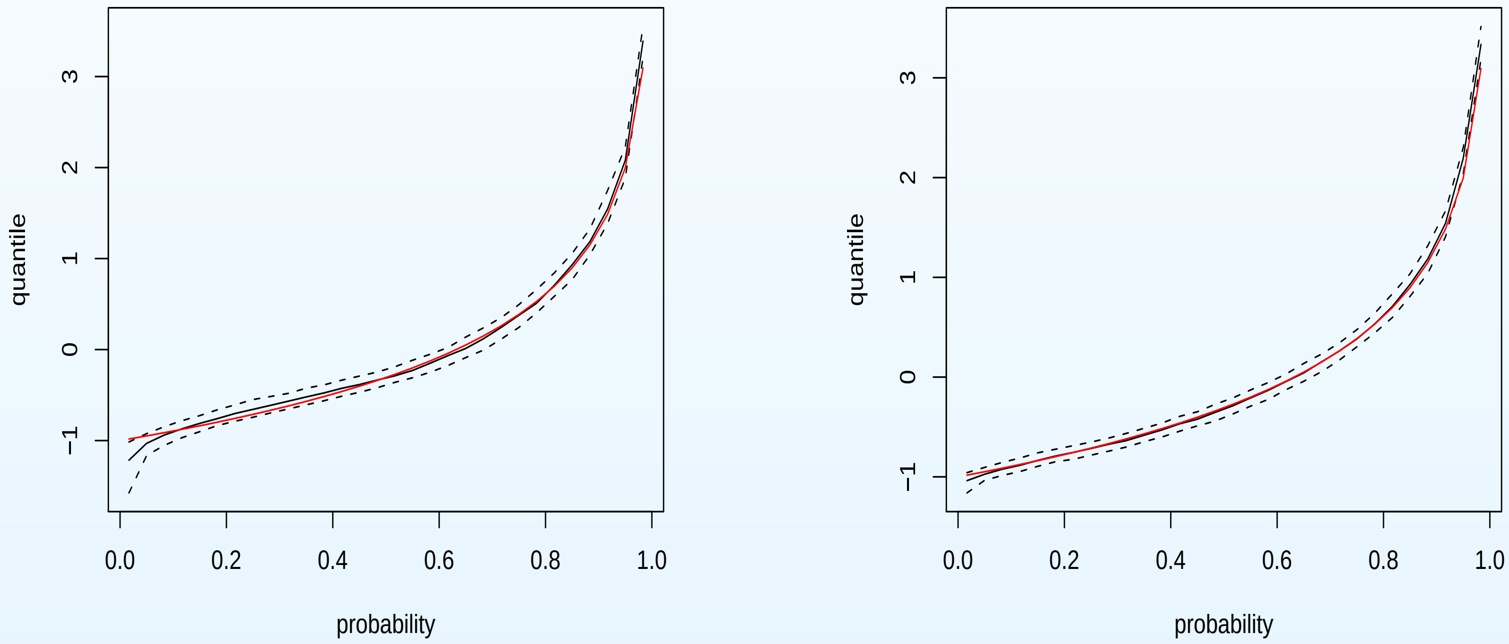


Red curve: F^{-1} Black curve: Median estimate of F^{-1}

Dashed curves: 10th and 90th percentiles of all estimates

$G =$ Normal, $F =$ Exponential

$\sigma = 1$ $\sigma = 3$

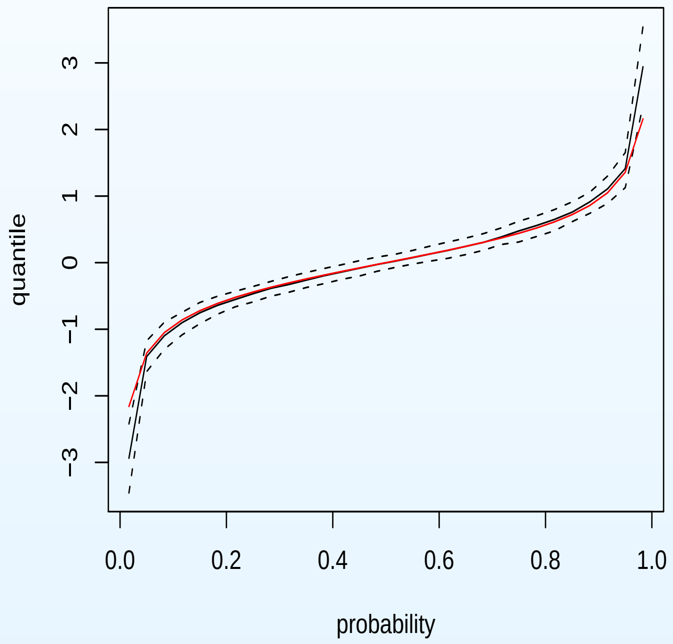


Red curve: F^{-1} Black curve: Median estimate of F^{-1}

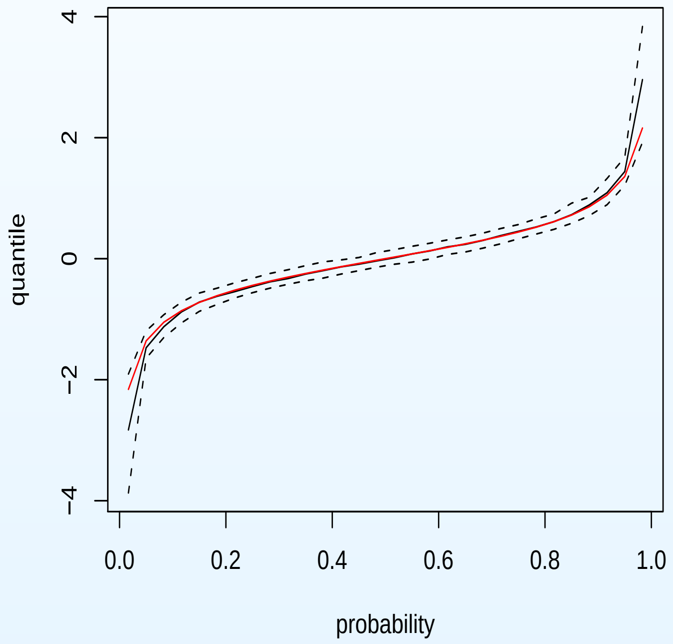
Dashed curves: 10th and 90th percentiles of all estimates

$G =$ **Normal,** $F = t_3$

$\sigma = 1$



$\sigma = 3$

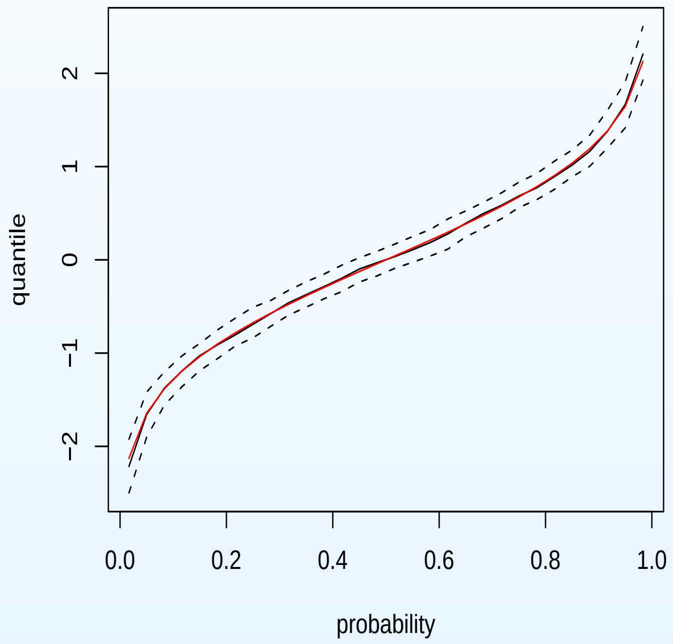


Red curve: F^{-1} Black curve: Median estimate of F^{-1}

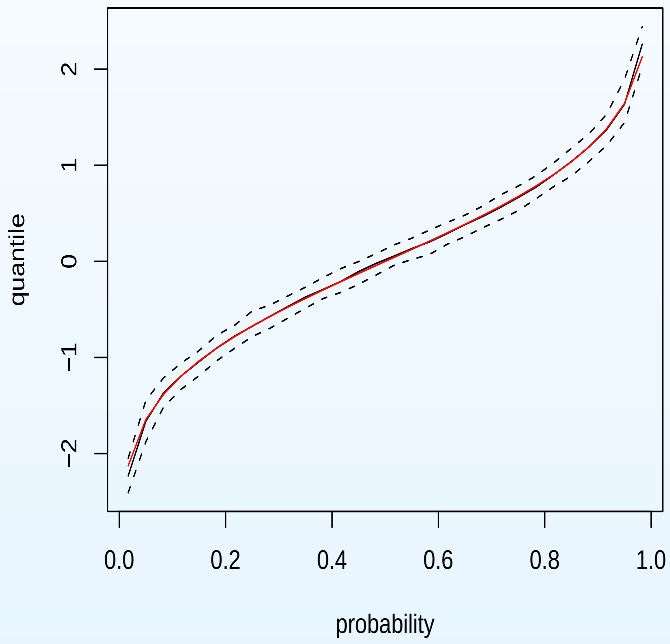
Dashed curves: 10th and 90th percentiles of all estimates

G = Normal mixture, F = Normal

$\sigma = 1$



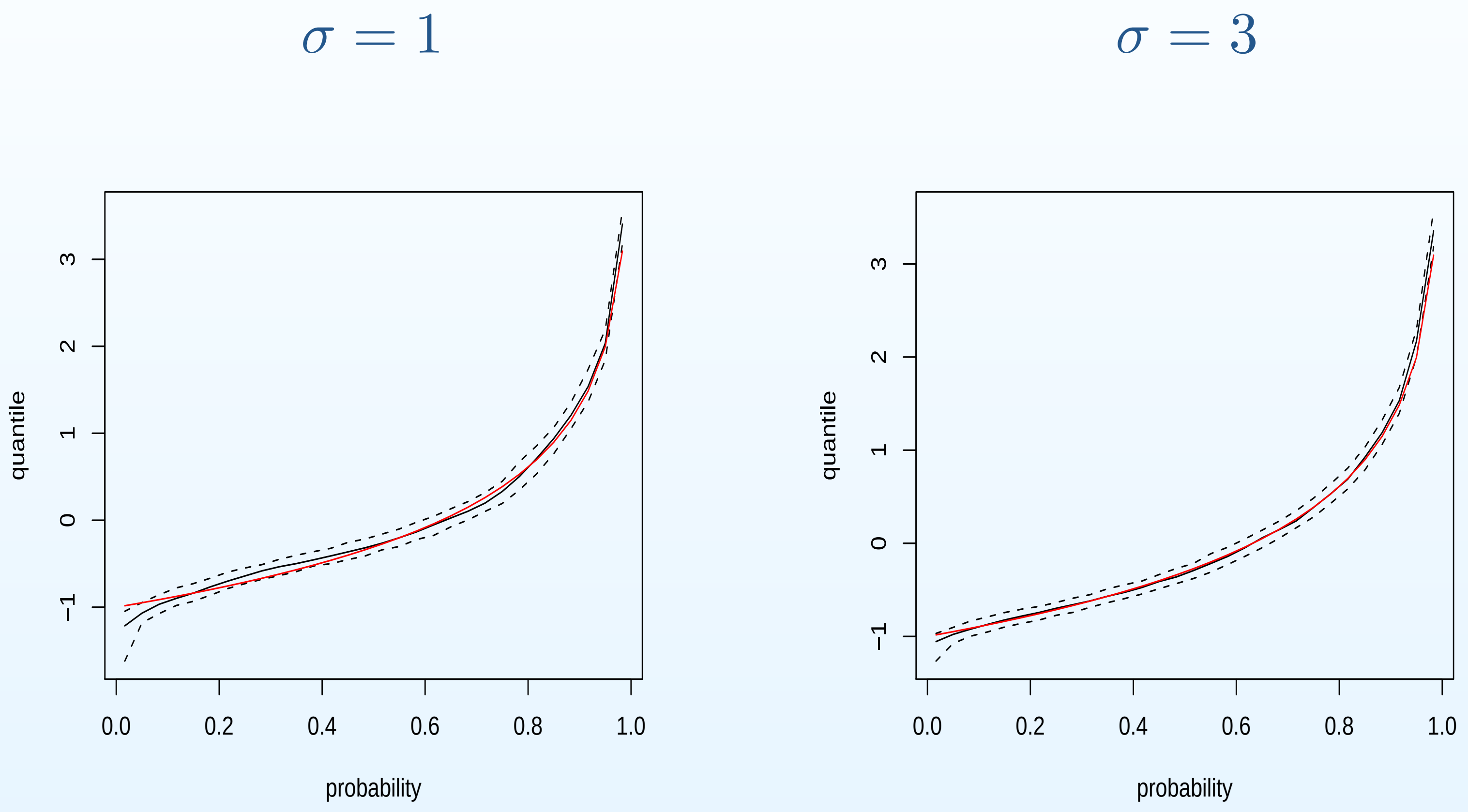
$\sigma = 3$



Red curve: F^{-1} Black curve: Median estimate of F^{-1}

Dashed curves: 10th and 90th percentiles of all estimates

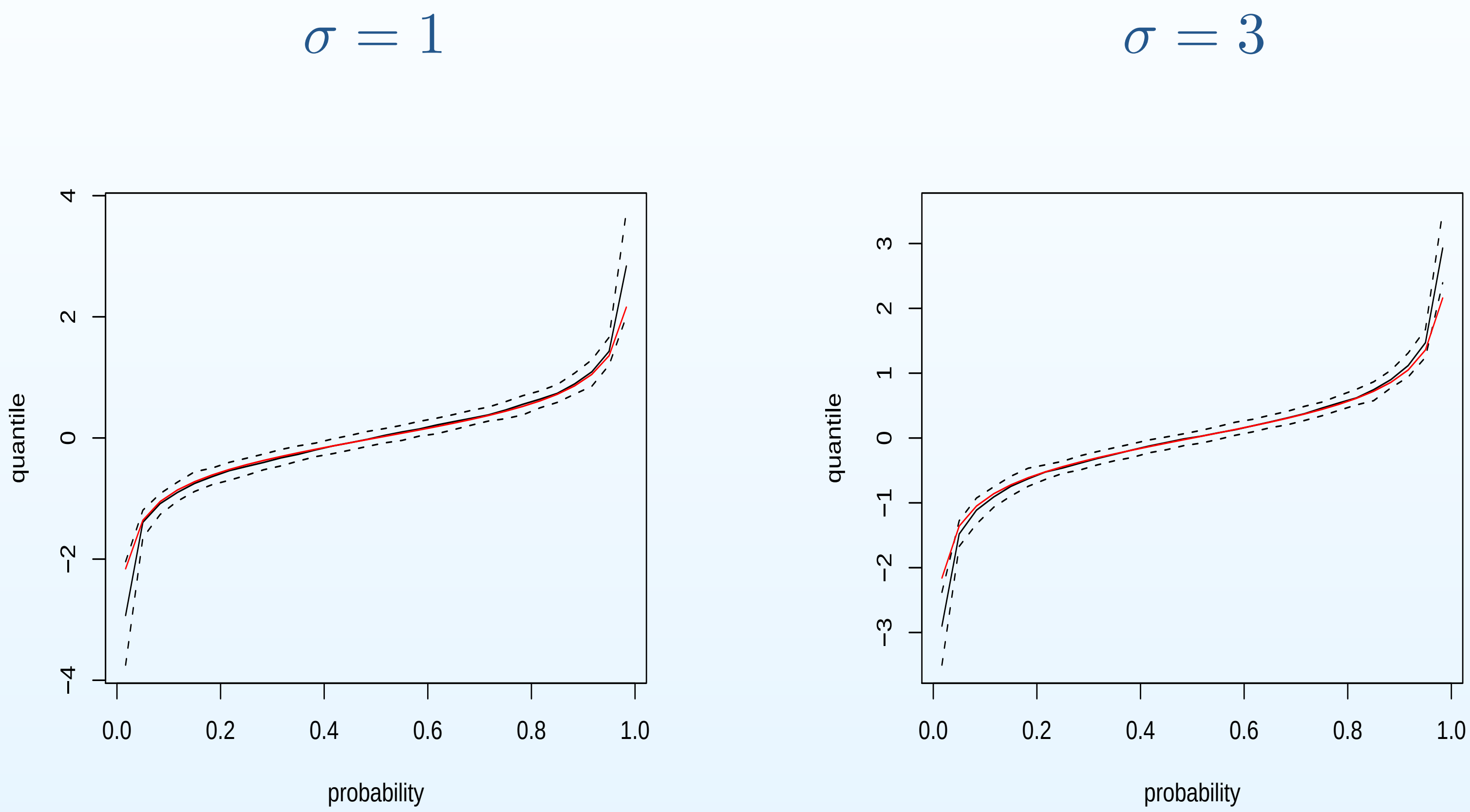
$G =$ Normal mixture, $F =$ Exponential



Red curve: F^{-1} Black curve: Median estimate of F^{-1}

Dashed curves: 10th and 90th percentiles of all estimates

$G =$ Normal mixture, $F = t_3$



Red curve: F^{-1} Black curve: Median estimate of F^{-1}

Dashed curves: 10th and 90th percentiles of all estimates

μ_i-free MDE estimates of F

Suppose $n \geq 3$ and define

$$\delta_{i1} = X_{i2} - X_{i1} = \sigma(\epsilon_{i2} - \epsilon_{i1})$$

and

$$\delta_{i2} = X_{i2} - X_{i3} = \sigma(\epsilon_{i2} - \epsilon_{i3}).$$

$(\delta_{i1}, \delta_{i2})$, $i = 1, \ldots, p$, are a special case of the location model.

It follows that the distribution of ϵ_{ij} is estimable from the differenced data!!

If one still wishes to estimate G, having a good estimate of F will help in this process.

Location-scale model

Suppose observed data are X_{ij}, $j = 1, \ldots, n$, $i = 1, \ldots, p$.

Consider the following model:

- $X_{ij} = \mu_i + \sigma_i \epsilon_{ij}$, $j = 1, \ldots, n$, $i = 1, \ldots, p$
- $(\mu_1, \sigma_1), \ldots, (\mu_p, \sigma_p)$ are i.i.d. as G_0.
- ϵ_{ij}, $j = 1, \ldots, n$, $i = 1, \ldots, p$, are i.i.d. as F_0, and independent of $(\mu_1, \sigma_1), \ldots, (\mu_p, \sigma_p)$.

MDE estimation of F based on residuals

In the location-scale model, the residuals

$$e_{ij} = \frac{X_{ij} - \bar{X}_i}{S_i} = \frac{\epsilon_{ij} - \bar{\epsilon}_i}{S_{\epsilon,i}}$$

are completely free of (μ_i, σ_i), $i = 1, \ldots, p$.

- f: Density of ϵ_{ij}.
- f_n: Corresponding density of e_{ij}.

Conjecture: Unless n is very small, f is identifiable from f_n.

MDE estimation from residuals, continued

Algorithm:

- Compute a kernel density estimate $\hat{f}_n$ of f_n from the residuals e_{ij}, $j = 1, \ldots, n$, $i = 1, \ldots, p$.
- Given a candidate $\tilde{f}$ for f, use simulation to approximate (arbitrarily well) the corresponding $\tilde{f}_n$.
- Compute $\int_{-\infty}^{\infty} (\hat{f}_n(x) - \tilde{f}_n(x))^2 \, dx$.
- Try to find a density $\tilde{f}$ such that the corresponding $\tilde{f}_n$ minimizes the distance in the previous step.

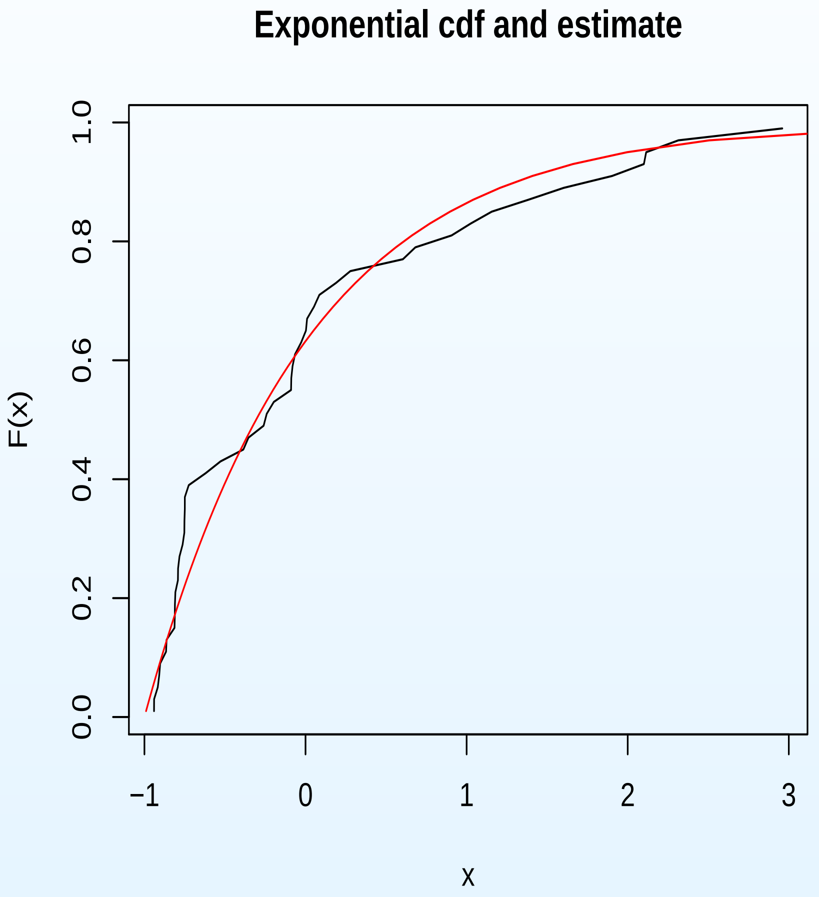
As candidate densities, use kernel smooths of candidate quantiles.

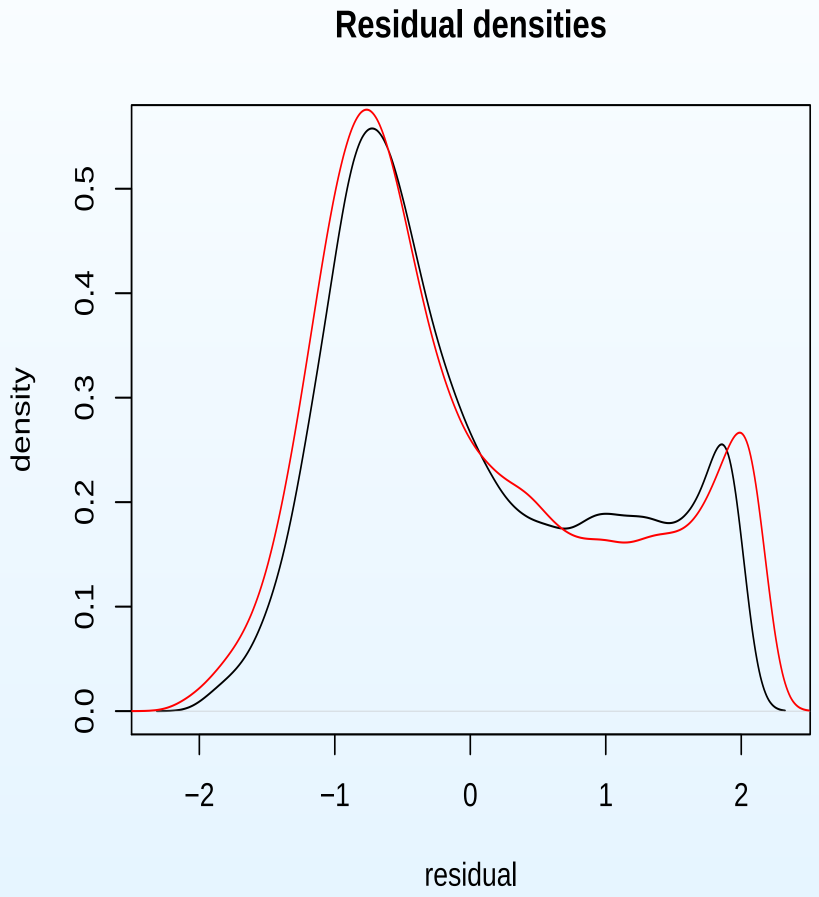
An example

Suppose $\epsilon_{ij} + 1$ has a standard exponential distribution.

- Generate $5(8038)$ values of ϵ_{ij}.
- Compute $5(8038)$ standardized residuals.
- Apply algorithm from previous page to estimate f.

Results for exponential example





Microarray example

Microarray data collected by Texas A&M nutritionist Robert Chapkin and coworkers.

The data here are a subset of data from a larger study.

- $n = 5$ rats that were all given the same treatment
- $p = 8038$ genes
- $X_{ij} = \log($expression level for gene j of rat $i)$

Model

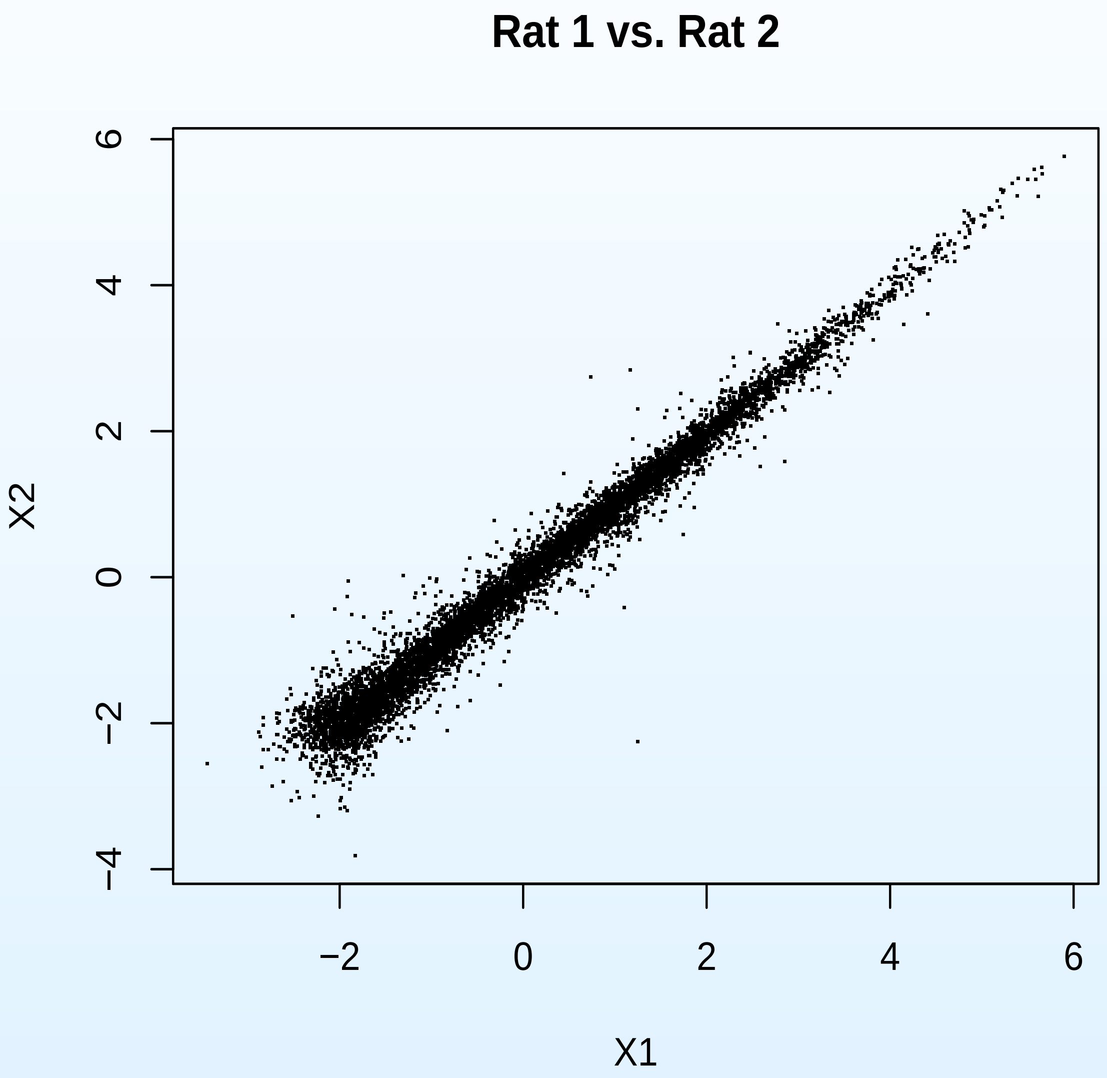
$$X_{ij} = M_j + \mu_i + \sigma_i \epsilon_{ij}, \quad j = 1, \ldots, 5, \; i = 1, \ldots, 8038.$$

- M_j: rat effect
- (μ_i, σ_i): gene effect
- ϵ_{ij}: error

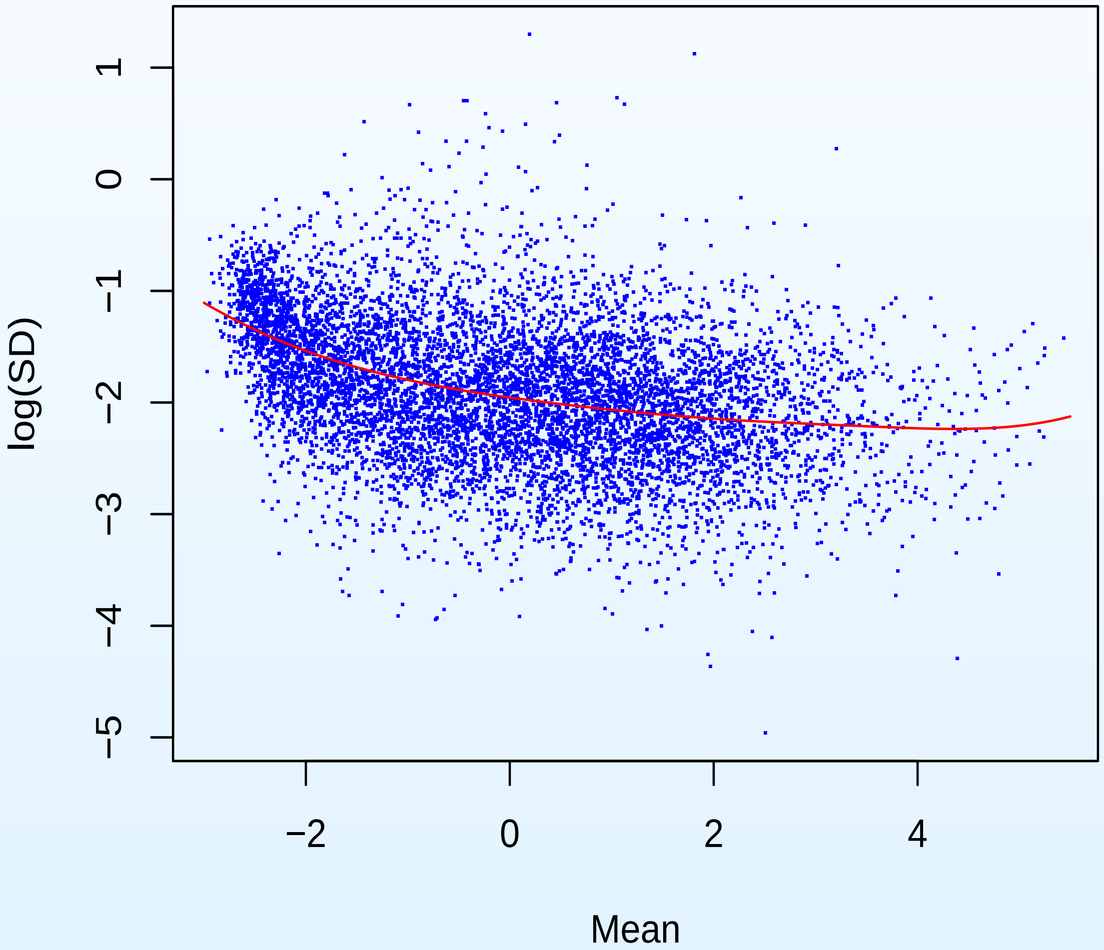
Remarks:

1. The rat effects can be very efficiently estimated since p is so large.
2. In this example our main interest is in estimating F, the distribution of each ϵ_{ij}.

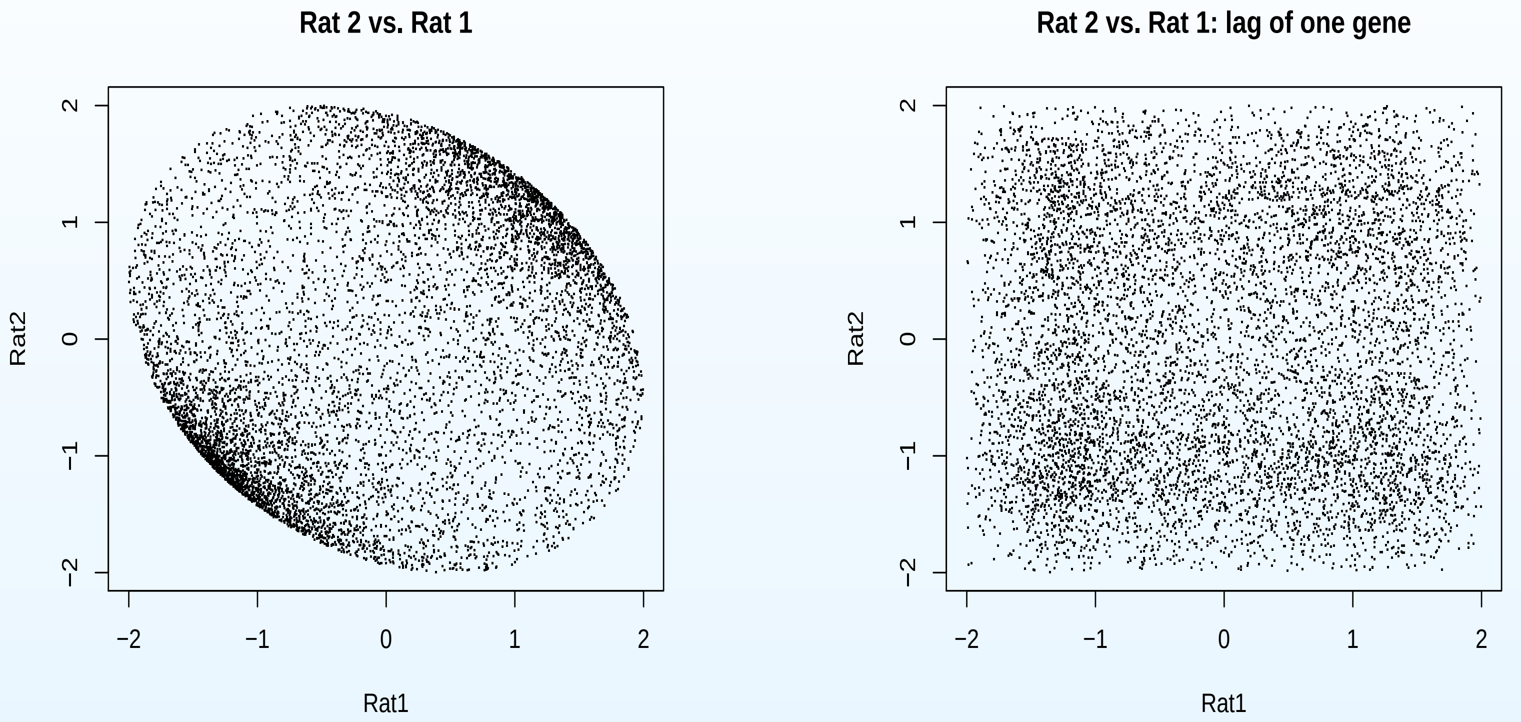
Typical scatterplot



Sample means and standard deviations

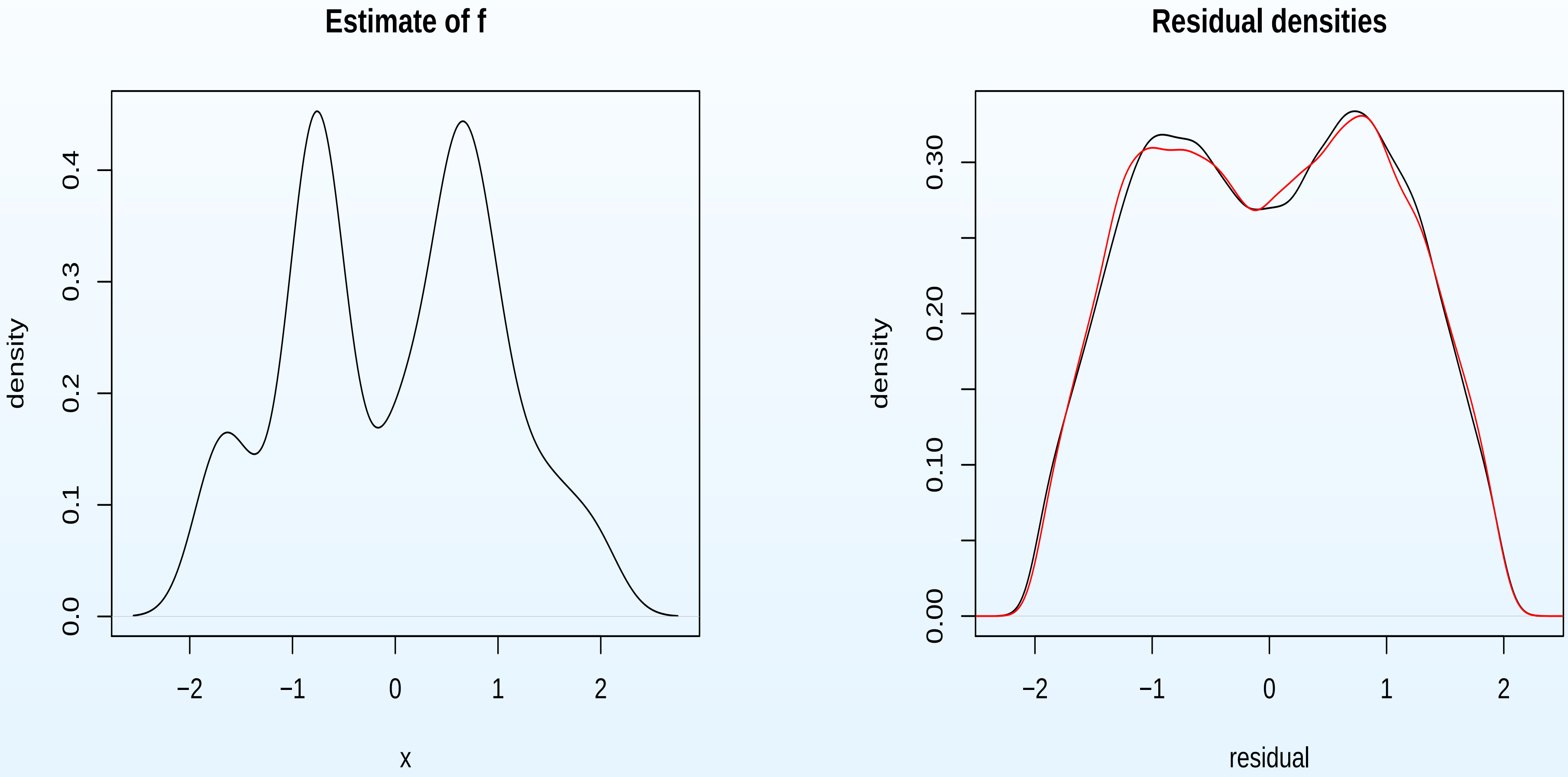


Residuals for two rats



- Elliptical pattern in the left plot is to be expected.
- Right plot is reassuring about independence between genes.

Density estimates



Right hand graph – Red: Kernel estimate of f_n Black: $\tilde{f}_n$

Further research

- Plenty of room to improve algorithm for approximating MDEs.
- Identifiability issues in location-scale model.
- Efficiency of MDE relative to explicit methods: ongoing work with Jan Johannes.