Dissertação apresentada à Faculdade de Ciências e Tecnologia da Universidade de Coimbra para obtenção do grau de Doutor em Engenharia Química na especialidade de Processos Químicos

Uma Comparação por Simulação de Monte Carlo de Estimadores de Regressão Não-Linear Robustos em Problemas de Engenharia Química

Eduardo Luis Trincão da Conceição

Coimbra, 2004

Departamento de Engenharia Química
Faculdade de Ciências e Tecnologia
UNIVERSIDADE DE COIMBRA

Dissertação realizada sob a orientação de

Prof. Doutor António Alberto Torres Garcia Portugal

e

Prof. Doutor José Almiro Abrantes de Menezes e Castro

Departamento de Engenharia Química da Universidade de Coimbra

À memória de meu pai, Carlos Luis da Conceição, e de meu professor, José Almiro Abrantes de Menezes e Castro.

Resumo

É sobejamente conhecido que o estudo de propriedades de amostra finita é de grande importância prática. Analisamos, neste trabalho, o desempenho de diversos estimadores de regressão não-linear *robustos* para modelos com uma única resposta e no caso de resposta múltipla, particularmente nos quadros de não normalidade do erro de medição e de contaminação da amostra por *outliers*. Com este propósito, realizaram-se experiências de Monte Carlo assentes em problemas reais essencialmente no domínio da cinética química. Seleccionámos para o caso univariado os métodos: L_p, mediana dos quadrados mínima (LMS), quadrados aparados mínimos (LTS), MM, τ, e diferenças aparadas mínimas (LTD). O caso multivariado inclui os seguintes métodos: M, mediana da verosimilhança máxima (MML), máxima verosimilhança aparada (MTL), e desvios absolutos mínimos (LAD). A estes juntámos os procedimentos mínimos quadrados (LS) e critério do determinante nas categorias univariada e multivariada, respectivamente, o que permite avaliar a competitividade dos métodos robustos relativamente a métodos clássicos. Em termos muito gerais, verifica-se a superioridade dos procedimentos robustos sobre os procedimentos clássicos. Visto que o número de problemas aqui analisados é pequeno, não será de estranhar que se tenha como essencial a realização de mais estudos de caso com vista a saber se a superioridade de um método sobre outro constatada no quadro de um problema pode generalizar-se ou não a outro problema. Relativamente ao caso univariado, os resultados dos diferentes estudos de caso aqui obtidos são coerentes entre si. O mesmo não se passa com o caso multivariado. No caso univariado o estimador MM destaca-se nos cenários sem *outliers* e de contaminação moderada, os quais constituem, em nossa opinião, as situações com que o analista se depara habitualmente. Nestas circunstâncias, os resultados sugerem a insensibilidade das estimativas MM ao tipo de estimador de alto ponto de rotura empregue no passo inicial.

Abstract

It is well known that the study of finite sample properties is of major practical importance. Here we report a study of the performance of several robust estimators for nonlinear regression both in single and multiresponse models. We used Monte Carlo simulation to study the influence of two main types of deviations from the usual assumptions, namely, lack of normality and contamination with outliers. The experiments were based in real life data taken from literature, mostly of chemical kinetics problems. The estimators analyzed for the univariate case are: L_p, mediana dos quadrados mínima (LMS), quadrados aparados mínimos (LTS), MM, τ, and diferenças aparadas mínimas (LTD). For the multivariate case we compare the following methods: M, mediana da verosimilhança máxima (MML), máxima verosimilhança aparada (MTL), and desvios absolutos mínimos (LAD). To assess the competitiveness of robust procedures over standard regression techniques, we also study the mínimos quadrados (LS) estimator for the univariate case and the determinant estimator for the multivariate case. Overall, the study has shown that the best robust estimators are indistinguishable or outperform the standard estimators. In view of the small number of problems, clearly more case studies are required to know whether the superiority of one method over another for one problem can be generalized to another problem. Our results reveal a general trend or behavior for the univariate case, but none for the multivariate case. The results for the univariate case indicate that the MM estimator is the most attractive technique for the two scenarios which we believe are the most important in practice, namely, when outliers are not present or in a moderate contamination scenario. Furthermore, the MM estimates seem to be insensitive to the type of high-breakdown initial estimator.

Agradecimentos

O meu agradecimento a F. Vegliò (Università degli Studi di Genova, Itália) e a Sergio Bobbo (Institute of Refrigeration, Itália) pelos úteis esclarecimentos sobre os casos de estudo da lixiviação de minério manganífero e do equilíbrio líquido-vapor da mistura refrigerante RE170 + R236fa, respectivamente.

Conteúdo

Lista de siglas e acrónimos

ARMENSI	**A**daptive **R**obust **M**-**E**stimator for **N**onparametric **S**ystem **I**dentification Estimador M Robusto Adaptativo para Identificação Não-Paramétrica de Sistemas
CSTR	**C**ontinuous **S**tirred **T**ank **R**eactor Reactor Tanque Contínuo com Agitação
DE	**D**ifferential **E**volution Evolução Diferencial
LAD	**L**east **A**bsolute **D**eviations Desvios Absolutos Mínimos
LMS	**L**east **M**edian of **S**quares Mediana dos Quadrados Mínima
LS	**L**east **S**quares Mínimos Quadrados
LTD	**L**east **T**rimmed **D**ifferences Diferenças Aparadas Mínimas
LTS	**L**east **T**rimmed **S**quares Quadrados Aparados Mínimos
MAD	**M**edian **A**bsolute **D**eviation Mediana dos Desvios Absolutos em Relação à Mediana
MCD	**M**inimum **C**ovariance **D**eterminant Determinante da Covariância Mínimo
MCRS	**M**odified **C**ontrolled **R**andom **S**earch Pesquisa Aleatória Controlada Modificada
MDE	**M**odified **D**ifferential **E**volution Evolução Diferencial Modificada
ML	**M**aximum **L**ikelihood Máxima Verosimilhança
MML	**M**aximum **M**edian **L**ikelihood Mediana da Verosimilhança Máxima

MSE	**M**ean **S**quared **E**rror Erro Quadrático Médio
MTL	**M**aximum **T**rimmed **L**ikelihood Máxima Verosimilhança Aparada
MVE	**M**inimum **V**olume **E**llipsoid Elipsóide de Volume Mínimo
PROGRESS	**P**rogram for **Ro**bust Re**gress**ion Programa para Regressão Robusta
WARME	**W**avelet-Based **A**daptive **R**obust **M**-**E**stimator Estimador M Robusto Adaptativo Baseado em Ôndulas

Lista de símbolos

Os números indicam a página em que o símbolo é introduzido ou definido. Os vectores são vectores coluna e serão denotados por letras minúsculas a negrito. As matrizes serão denotadas por letras maiúsculas a negrito.

A factor de frequência ou pré-exponencial (48)

a factor de consistência para a estimativa de escala na proposta 2 de Huber dos estimadores M (19)

a parâmetro da equação de estado de Carnahan-Starling-De Santis (52)

a ordem parcial de reacção em relação ao ácido sulfúrico (58)

b enviesamento máximo das estimativas causado pela contaminação da amostra (5)

b constante que forma o 2.º membro da equação de definição de s_n (22)

b parâmetro da equação de estado de Carnahan-Starling-De Santis (52)

b_1, b_2 parâmetros que relacionam a energia de activação com a conversão (58)

C constante do modelo de lixiviação de minério manganífero (58)

C constante do modelo de crescimento de células dependentes de ancoragem (62)

$\boldsymbol{C}$ estimativa de dispersão do estimador determinante da covariância mínimo (28)

c concentração (48)

c ordem parcial de reacção em relação ao hidrato de carbono (58)

C_{R} factor de *crossover* do algoritmo MDE (33)

D^m conjunto da totalidade de permutações possíveis de contaminação duma amostra (5)

E esperança matemática (11)

E energia de activação (48)

F função de distribuição principal (4)

f modelo matemático mecanístico do sistema ou processo (3)

$\boldsymbol{f}$ vector de funções que constituem o modelo matemático do processo (26)

f_0 constante cujo valor reflecte a variabilidade da variável dependente nas observações (35)

F_{b} factor de ponderação base do algoritmo MDE (33)

G função de distribuição das observações ou do erro de medição (4)

g função densidade de probabilidade genérica (13)

$\Delta H_k^{\mathrm{fus}}$ entalpia molar de fusão do componente k (55)

H função de distribuição contaminante (4)

h cobertura (20)

$\mathcal{H}$ constante da lei de Henry (124)

K constante de adsorção (45)

k_{12} parâmetro de interacção binária (52)

k_2 constante de velocidade para a segunda etapa da hidrogenação do tolueno, $k_2 = k_{\mathrm{H},2} + k_{\mathrm{D},2}$ (113)

k_{a} constante de velocidade para a adsorção de oxigénio (48)

k_{D} constante de velocidade de disproporcionação (113)

k_{H} constante de velocidade de hidrogenação (113)

K_k coeficiente de partição líquido-vapor do componente k (54)

K_p constante de equilíbrio, na base da pressão (45)

k_{r} constante de velocidade para a oxidação do propeno (48)

ℓ componente da função log-verosimilhança (26)

L função de verosimilhança (14)

l função usada na definição de R_n (6)

l logaritmo da função de verosimilhança (14)

MD distância de Mahalanobis (183)

m número de pontos da amostra original contaminados (5)

N número de réplicas de Monte Carlo (3)

n número de observações, tamanho ou dimensão da amostra (3)

n número estequiométrico (número de moles de oxigénio consumidas por mole de propeno oxidado) (48)

n_c número de observações regulares (40)

n_o número de *outliers* (40)

n_p dimensão do vector de parâmetros do modelo (3)

N_{P} tamanho da população do algoritmo MDE (33)

n_t dimensão do vector de variáveis de decisão do algoritmo MDE (33)

n_x dimensão do vector das variáveis independentes (3)

n_y dimensão do vector de variáveis dependentes (26)

p parâmetro de forma da função densidade da distribuição exponencial-potência (13)

p pressão (45)

R constante dos gases (48)

r' razão da velocidade de reacção pela massa de catalisador (45)

RB estimativa robusta de enviesamento usada como índice de desempenho de um estimador numa experiência de Monte Carlo (44)

RD distância robusta de Rousseeuw e van Zomeren (183)

R_n critério de medida do grau de deficiência duma estimativa (6)

R_p tamanho médio das partículas de minério manganífero (58)

S função objectivo do 3.º passo do estimador MM (22)

s estimativa de dispersão do erro de medição para o estimador dos mínimos quadrados (12)

s_n estimativa de dispersão do erro de medição para os estimadores MM e τ (22)

T estimador de regressão (4)

T temperatura (48)

t probabilidade de contaminação (4)

t tempo (58)

t_0 "tempo morto" (117)

T' temperatura de referência (58)

$T_{m,k}$ temperatura de fusão do componente k (55)

V volume molar (52)

$\boldsymbol{V}$ matriz de covariâncias das variáveis independentes (183)

X grau de conversão (58)

x densidade média de células na cultura (62)

x_∞ densidade máxima de células que pode ser atingida em confluência (62)

$\boldsymbol{x}$ vector de variáveis independentes ou explicativas (3)

x_k fracção molar do componente k na fase líquida (54)

y variável dependente ou resposta (3)

$\boldsymbol{y}$ vector de variáveis dependentes (26)

y_k fracção molar do componente k na fase de vapor (54)

z amostra (4)

z_k fracção molar do componente k (52)

z^m amostra contaminada (5)

Letras gregas

α fracção de aparamento (20)

α logaritmo da constante de velocidade à temperatura T_1 (48)

β logaritmo da constante de velocidade à temperatura T_2 (48)

Γ função Gama (13)

γ parâmetros perturbadores (34)

γ_k coeficiente de actividade do componente k (55)

Δ_x distribuição que coloca a totalidade da sua massa no ponto x (5)

$\boldsymbol{\Delta_y}$ desvio em y (40)

δ_k parâmetro de solubilidade do componente k (56)

δ_{R} número de desvios padrão do erro que os *outliers* distam dos pontos regulares no espaço-y (40)

ϵ erro ou ruído de medição (3)

$\boldsymbol{\epsilon}$ vector dos erros de medição (26)

ϵ_n^* ponto de rotura dum estimador (5)

ϵ_n^+ função de rotura superior de Stromberg e Ruppert (6)

ϵ_n^- função de rotura inferior de Stromberg e Ruppert (6)

ϵ_n função de rotura de Stromberg e Ruppert (6)

$\tilde{\epsilon}_n$ ponto de rotura de Stromberg e Ruppert (6)

ε_0 tolerância do critério de convergência dos algoritmos evolutivos (35)

θ fracção da superfície que está coberta por moléculas adsorvidas (113)

$\boldsymbol{\theta}$ vector de parâmetros do modelo matemático (3)

$\hat{\boldsymbol{\theta}}$ vector das estimativas dos parâmetros (12)

λ_{12} constante de interacção binária de uma extensão da teoria das soluções regulares (56)

μ parâmetro de localização da função densidade da distribuição exponencial-potência (13)

$\mu_{\max}$ taxa específica máxima de crescimento (62)

$\boldsymbol{\mu}$ média aritmética das variáveis independentes (183)

ν coeficientes estequiométricos (58)

ξ passo do algoritmo MDE (33)

ρ função que integra a definição dos estimadores M (16)

σ factor de escala usado para tornar os estimadores M equivariantes à escala de y (19)

σ_ϵ^2 variância do erro de medição (12)

$\boldsymbol{\Sigma}$ matriz de covariâncias das variáveis dependentes (26)

ϕ parâmetro de escala da função densidade da distribuição exponencial-potência (13)

Φ_k fracção volúmica do componente k (56)

ϕ_k coeficiente de fugacidade do componente k (54)

Índices

calc valor calculado (53)

exp valor medido (53)

i denota observações (3)

$i:n$ i-ésima estatística de ordem duma colecção de n números (20)

i_c denota observações regulares (40)

i_o denota *outliers* nos dados (40)

j denota respostas (26)

k denota componentes (52)

Expoentes

0 inicial (58)

$\hat{}$ valor estimado (12)

L limite inferior (7)

L fase líquida (54)

rel valor relativo (113)

U limite superior (7)

V fase de vapor (54)

Capítulo 1

Introdução geral

Este capítulo apresenta o tema da investigação (o estudo empírico do comportamento de estimadores robustos) e sua motivação, assim como descreve o instrumento usado para esse estudo (o método de Monte Carlo). Além disso, descreve algumas noções básicas de estimação robusta e analisa o problema de inferência estatística nesse contexto.

1.1 Tema e motivação da investigação

O tema deste trabalho é o estudo empírico do desempenho de um conjunto de técnicas de estimação robusta no âmbito da Engenharia Química.

Em termos muito gerais, o objectivo dos procedimentos estatísticos robustos é a insensibilidade a pequenos desvios dos pressupostos dos métodos clássicos (Huber, 1996, p. 1). Como decorre da variedade de hipóteses ordinariamente usadas em estatística, o conceito de robustez é demasiado abrangente para que seja praticável o seu estudo sem restringir o conjunto de pressupostos considerado, originando-se assim, diferentes campos parcelares da robustez. No contexto da regressão ou estimação de parâmetros, a área mais desenvolvida é a que estuda a resistência dos estimadores a desvios do modelo Gaussiano do erro ou ruído de medição, bem como à presença de *outliers* na amostra.

A literatura estatística de regressão robusta investiga fundamentalmente o modelo *linear* de reposta univariada, apresentando uma multiplicidade de estimadores e procedimentos alternativos ao estimadores clássicos de máxima verosimilhança e dos mínimos quadrados. Na sua maioria, a forma de definição das respectivas estimativas — basicamente a determinação de um extremo duma função dos resíduos —, permite a sua extensão directa ao caso não-linear (Stromberg e Ruppert, 1992). Uma questão que se põe é a seguinte: dada a diversidade de estimadores disponíveis, qual é a qualidade do desempenho de determinado estimador?

Infelizmente, a análise de conceitos de robustez no âmbito da regressão não-linear é escassa (Müller, 1997, p. 186; Ryan, 1997, p. 436) — uma pesquisa na base de dados bibliográficos Zentralblatt MATH conduziu a um total de 35 artigos —, e consequentemente para inúmeros estimadores os estudos ou não existem ou são incompletos. Por outro lado, os poucos resultados teóricos disponíveis são assimptóticos ou de *grande amostra*, isto é, válidos para uma situação limite da dimensão da amostra.

Ora, na realidade as amostras são de dimensão finita, situação para a qual a teoria assimptótica clássica não fornece respostas. Na prática corrente, os resultados assimptóticos são interpretados como uma aproximação para amostras finitas, sem um critério que permita estabelecer de maneira rigorosa o valor do tamanho da amostra a partir do qual essa aproximação é suficientemente precisa (Hampel, 1998). Por outras palavras,

as propriedades assimptóticas apenas se observam *provavelmente* em amostras muito grandes. Le Cam e Yang (2000, pp. 175–177) criticam de igual forma a ênfase existente na teoria assimptótica:

> It must be pointed out that the asymptotics of the "standard i.i.d. case" are of little relevance to practical use of statistics, in spite of their widespread study and use. ... The use of such considerations is an abuse of confidence that has been foisted upon unsuspecting students and practitioners owing to the fact that we, as a group, possess limited analytical abilities and, perforce, have to limit ourselves to simple problems. ...
>
> Common textbooks often include statements about "consistency" of estimates. Apparently, it is considered good for estimates $\hat{\theta}_n$ to converge in probability to the "true value" θ as $n \to \infty$. It is even better if they do so almost surely. Few texts point out that such a property is entirely irrelevant to anything of value in practice. ...
>
> The use of asymptotics "as $n \to \infty$" for the standard i.i.d. case seems to be based on an entirely unwarranted act of faith. If you do prove that $\hat{\theta}_n$ has some good asymptotic property, maybe some of that percolate down to $n = 10^6$ or even $n = 100$.

Tendo em conta que devido a restrições de natureza tecnológica e económica as amostras em experimentação planeada são tipicamente de pequena dimensão, o uso de propriedades assimptóticas é questionável. È necessário, portanto, atacar directamente o problema de avaliação da qualidade de um estimador no contexto de amostras de pequena dimensão. Que vias é possível seguir?

Em Field e Ronchetti (1990) é descrita uma abordagem teórica com a qual é possível obter aproximações assimptóticas precisas para amostras de dimensão extremamente pequena (eventualmente até $n = 1$!). Infelizmente, não se conhecem aplicações no contexto da regressão robusta. Assim, a abordagem padrão de ataque a este problema é a simulação ou experimentação de Monte Carlo (Huber, 1996, p. 62; Müller, 1997, p. 184; Hampel, 1998). A comparação e caracterização do desempenho de diferentes estimadores faz-se essencialmente pelo estudo do enviesamento e variância amostrais obtidos, os quais podem ser combinados num critério como o erro quadrático médio.

1.2 Algumas noções básicas sobre simulação de Monte Carlo

Um estudo de Monte Carlo pode ser utilizado com três propósitos (Gentle, 1998, pp. 178 e 179):

1. efectuar a avaliação preliminar das características de um procedimento estatístico,
2. determinar as propriedades de um método analiticamente intratável, e,
3. efectuar o estudo das propriedades em amostras finitas.

Os elementos básicos no método de Monte Carlo serão, naturalmente, explicados no contexto do modelo de regressão, que é objecto deste trabalho. Consideremos, então,

a título de exemplo, o modelo estatístico usual de regressão não-linear com resposta univariada, no qual se dispõe de n observações $(\boldsymbol{x}_i, y_i)$:

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\theta}) + \epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{1.1}$$

onde y é a variável dependente, $\boldsymbol{x}$ é o vector $n_x \times 1$ de variáveis independentes, $\boldsymbol{\theta}$ é o vector $n_p \times 1$ de parâmetros do modelo matemático, f representa o modelo matemático mecanístico do sistema, e ϵ é o erro ou ruído de medição.

Uma experiência "do mundo real" consiste em medir a variável dependente para valores predeterminados de $\boldsymbol{x}$ mediante um procedimento estatístico formal de planeamento de experiências ou por intuição do experimentalista.

Um estudo de Monte Carlo consiste em realizar um grande número, N, de réplicas de uma experiência artificial para cada combinação de condições que se pretendem investigar. Esta experiência efectua-se especificando um modelo "verdadeiro" — o que no contexto do modelo (1.1) significa especificar as variáveis independentes, os valores dos parâmetros, a distribuição do erro aleatório, e o tamanho da amostra —, e gerar um conjunto de observações (os valores de y_i). O critério usado para medir o desempenho do estimador é calculado a partir das estimativas dos parâmetros obtidas pela aplicação do estimador aos dados de cada replicação. Trata-se pois de um procedimento empírico, que, como tal, apresenta a fraqueza de conduzir a conclusões específicas do caso especial em estudo (Johnston e DiNardo, 2001, p. 389).

Uma das linhas de orientação para o delineamento de experiências de Monte Carlo é a de que o mundo de Monte Carlo deve imitar o mundo real de forma a que os resultados sejam relevantes "para a compreensão de problemas com dados reais" (Johnston e DiNardo, 2001, p. 382). Por outras palavras, as condições experimentais com que o experimentalista se confronta no mundo real devem ser transpostas na medida do possível para o estudo de Monte Carlo. A forma de alcançar este objectivo de forma pragmática consiste numa abordagem de estudo de caso de problemas *reais* de estimação de parâmetros recolhidos da literatura de engenharia química.

Os factores investigados nas experiências de Monte Carlo apresentadas neste trabalho incluem:

- a distribuição do componente aleatório do modelo (ruído de medição),
- a contaminação da amostra por *outliers*, e
- a classe do estimador e variantes.

Deste modo, o resultado final da experiência de Monte Carlo pode ser expresso da seguinte forma:

$$C(\hat{\boldsymbol{\theta}}_{ijk}) = h(\text{distribuição}_i, \text{estimador}_j, \text{contaminação}_k),$$

onde

$$\hat{\boldsymbol{\theta}}_{ijk} = \begin{bmatrix} \theta_{11} & \theta_{12} & \ldots & \theta_{1n_p} \\ \theta_{21} & \theta_{22} & \ldots & \theta_{2n_p} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{N1} & \theta_{N2} & \ldots & \theta_{Nn_p} \end{bmatrix}$$

Inicialização: Estipular os "verdadeiros" valores dos parâmetros do modelo matemático e dos parâmetros das diferentes distribuições arbitradas para o erro de medição, de acordo com os dados da experiência real.
Fixar o tamanho da amostra de cada experiência simulada.
para cada amostra **fazer** (Ciclo da dimensão amostral)
  Fixar os valores das variáveis independentes (planeamento de experiências), ou usar os valores da experiência real.
  para cada distribuição seguida pelo erro de medição **fazer** (Ciclo das distribuições)
    para um dado número de réplicas **fazer** (Ciclo de Monte Carlo)
      Gerar um conjunto de observações para o modelo de regressão não-linear
      para cada estimador **fazer** (Ciclo dos estimadores)
        Proceder ao cálculo da estimativas dos parâmetros do modelo mecanístico.
      fim (para)
    fim (para)
  fim (para)
fim (para)
Calcular os critérios de desempenho de cada estimador para os diferentes parâmetros do modelo matemático.

Figura 1.1 Estrutura de uma experiência de Monte Carlo (adaptado de Gentle, 1998, p. 189).

é a matriz $N \times n_p$ de estimativas dos parâmetros do modelo mecanístico da (ijk)-ésima combinação dos níveis dos factores (as linhas desta matriz são as diversas réplicas das estimativas dos n_p parâmetros, e cada coluna corresponde às N réplicas de um parâmetro), e C representa um critério de desempenho, calculado para cada parâmetro a partir da respectiva coluna de $\hat{\boldsymbol{\theta}}_{ijk}$.

A implementação computacional duma experiência de Monte Carlo traduz-se essencialmente por um conjunto de ciclos sucessivamente encaixados, os quais varrem os níveis de cada factor da experiência. O cálculo principal corresponde à determinação das estimativas dos parâmetros. A estrutura geral de um programa para uma experiência de Monte Carlo é apresentada nesta página.

1.3 Definições de robustez

Nesta secção apresentam-se duas formas de avaliação da robustez de um estimador, o *ponto de rotura* e a *função de influência*.

Designe-se por G a distribuição das observações e represente-se por $T(G)$ um estimador de regressão. Uma lei de erro frequentemente usada em estatística robusta para modelar as amostras que ocorrem no campo das ciências exactas e de engenharia é a mistura de distribuições

$$G = (1 - t)F + tH,$$

que produz observações com elevada probabilidade $1 - t$ da distribuição principal F contaminadas por "más" observações da distribuição H com pequena probabilidade t. A geração de uma amostra $z \sim G$ pode encarar-se como um procedimento em dois passos:

no primeiro selecciona-se a distribuição F ou H com as probabilidades respectivas $1-t$ e t. Em seguida gera-se $z \sim F$ ou $z \sim H$ a partir da distribuição escolhida. Numa amostra de dimensão elevada, a fracção de observações contaminantes (ou contaminação) é aproximadamente t.

Se tomarmos para H a distribuição que coloca a totalidade da sua massa no ponto x, Δ_x, então para um valor "elevado" de x, $G = (1-t)F + t\Delta_x$ é um modelo simplificado para uma situação de contaminação duma amostra por uma proporção t de *outliers* em x.

A função de influência (IF) de um estimador T é dada por

$$\text{IF} = \lim_{t \to 0} \frac{T((1-t)F + t\Delta_x) - T(F)}{t} \tag{1.2}$$

nos pontos x em que o limite existe, e mede o efeito que uma pequena fracção de contaminação por *outliers* no ponto x provoca no estimador.

O conceito de ponto de rotura de amostra finita de um estimador introduzido por Donoho e Huber (1983) define-se informalmente como a fracção mínima de contaminação da amostra por *outliers* que provoca a "inutilização" da estimativa. A formalização desta "inutilização" depende do contexto considerado. No contexto da regressão linear, significa uma variação ilimitada no valor da estimativa.

Mais precisamente, seja $z = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ uma amostra com n observações e T um estimador de regressão, o qual aplicado à amostra conduz às estimativas dos parâmetros, isto é, $T(z) = \hat{\theta}$. Represente-se por D^m o conjunto de todas as amostras contaminadas z^m possíveis obtidas a partir de z pela substituição de m dos pontos originais por valores arbitrários, e por $b(n, m; T, z)$ o enviesamento máximo causado por tal contaminação, ou seja,

$$b(n, m; T, z) = \sup_{z^m \in D^m} d(T(z), T(z^m)),$$

onde d é uma medida de distância, por exemplo, $d(T(z), T(z^m)) = \|T(z^m) - T(z)\|$.

Então o ponto de rotura do estimador T na amostra z é definido por

$$\epsilon_n^*(T, z) = \min_{0 \leqslant m \leqslant n} \left(\left\{ \frac{m}{n} : b(n, m; T, z) = \infty, \;\; m \in \{1, \ldots, n\} \right\} \cup \{1\} \right). \tag{1.3}$$

Observe-se que a operação de união na expressão de ϵ_n^* está lá para assegurar que o conjunto empregado para a sua definição não seja vazio.

Stromberg e Ruppert (1992) notam duas deficiências na definição anterior quando aplicada no âmbito da regressão não-linear:

1. se o espaço dos parâmetros for limitado — o que acontece tipicamente em modelos mecanísticos em que os parâmetros possuem significado físico —, então, pela sua definição, $\epsilon_n^* = 1$, pois d não pode tomar valores infinitamente grandes e, por isso, o subconjunto da esquerda em (1.3) é vazio;

2. por outro lado, o valor do ponto de rotura não é invariante à reparametrização de $\boldsymbol{\theta}$ no modelo matemático.

De modo a ultrapassar as deficiências apontadas, os autores propõem uma definição baseada no comportamento do modelo estimado $f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ no lugar do de $\hat{\boldsymbol{\theta}}$. Mais precisamente, definem a função de rotura superior, ϵ_n^+, a função de rotura inferior, ϵ_n^-, e a função de rotura ϵ_n no ponto $\boldsymbol{x}$, para o modelo f, estimador T, e amostra z, por

$$\tilde{\epsilon}_n^+(\boldsymbol{x}, f, T, z) = \begin{cases} \min\limits_{0 \leqslant m \leqslant n} \left\{ \dfrac{m}{n} : \sup\limits_{z^m \in D^m} f(\boldsymbol{x}, T(z^m)) \right. \\ \qquad\qquad \left. = \sup\limits_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta}), \;\; m \in \{1, \ldots, n\} \right\} & \text{se } \sup_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta}) > f(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) \\ 1 & \text{caso contrário,} \end{cases} \tag{1.4}$$

$$\tilde{\epsilon}_n^-(\boldsymbol{x}, f, T, z) = \begin{cases} \min\limits_{0 \leqslant m \leqslant n} \left\{ \dfrac{m}{n} : \sup\limits_{z^m \in D^m} f(\boldsymbol{x}, T(z^m)) \right. \\ \qquad\qquad \left. = \inf\limits_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta}), \;\; m \in \{1, \ldots, n\} \right\} & \text{se } \inf_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta}) < f(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) \\ 1 & \text{caso contrário,} \end{cases} \tag{1.5}$$

e

$$\tilde{\epsilon}_n(\boldsymbol{x}, f, T, z) = \min\{\tilde{\epsilon}_n^-(\boldsymbol{x}, f, T, z), \tilde{\epsilon}_n^+(\boldsymbol{x}, f, T, z)\}. \tag{1.6}$$

Por palavras, os valores das funções de rotura superior e inferior definem-se para cada ponto $\boldsymbol{x}$ e estimador T como a proporção mínima de contaminação que conduz o modelo a $\sup_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta})$ e $\inf_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta})$, respectivamente, enquanto se designa por função de rotura a fracção mínima de contaminação que provoca a rotura superior ou inferior do estimador.

Por último, define-se o ponto de rotura, $\tilde{\epsilon}_n$, por

$$\tilde{\epsilon}_n(f, T, z) = \inf_{\boldsymbol{x}}\{\epsilon(\boldsymbol{x}, f, T, z)\} \tag{1.7}$$

Os autores apontam a possibilidade de nalgumas circunstâncias o ponto de rotura assim definido não ser apropriado, advogando o uso de um ponto de rotura superior ou inferior definido de forma similar.

Por sua vez, Sakata e White (1995) notam que nem sempre a convergência do modelo para os seus valores extremos constitui um critério adequado de medida do grau de deficiência duma estimativa. Por conseguinte apresentam uma definição alternativa para ponto de rotura, conquanto considerem que esta não é necessariamente uniformemente superior à de Stromberg e Ruppert.

A ideia básica é que cabe ao utilizador a definição do critério de medida do grau de deficiência da estimativa. De novo, de modo a garantir a invariância a reparametrizações do modelo, este deverá tomar em atenção o comportamento do modelo estimado, de preferência ao comportamento das estimativas dos parâmetros. Uma possível forma é

$$R_n(f(\boldsymbol{x}, T(z)), z) = n^{-1} \sum_{i=1}^{n} l[y_i - f(\boldsymbol{x}, T(z))],$$

onde l é uma função decrescente em $(-\infty, 0]$ e crescente em $[0, \infty)$. Formalmente, o ponto de rotura baseado numa medida de deficiência pode ser escrito numa versão simplificada

como

$$\min_{0\leqslant m\leqslant n}\left(\left\{\frac{m}{n}:\sup_{z^m\in D^m}R_n(f(\boldsymbol{x},T(z^m)),z)=\sup_{\boldsymbol{\theta}}R_n(f(\boldsymbol{x},\boldsymbol{\theta}),z),\right.\right.$$
$$\left.\left.m\in\{1,\ldots,n\}\right\}\cup\{1\}\right). \quad (1.8)$$

Um aspecto saliente nas definições de Stromberg e Ruppert e Sakata e White é a invariância a reparametrizações do modelo. Contudo, quando se utilizam modelos mecanísticos existe uma parametrização "natural" que decorre do significado físico dos parâmetros. Neste quadro, o uso da reparametrização num procedimento de estimação de parâmetros coloca-se em duas circunstâncias. Em primeiro lugar, com a finalidade de melhorar o desempenho do algoritmo usado, sendo as estimativas obtidas convertidas de novo para a forma original. Em segundo lugar, pode acontecer que a qualidade das propriedades estatísticas das estimativas de determinados parâmetros seja pobre. A reformulação destes parâmetros pode constituir a única forma de contornar este problema, mas pagando o preço da perda do significado físico original. A questão que se põe é decidir entre o uso dos parâmetros originais com más propriedades estatísticas, ou de parâmetros não convencionais bem comportados (Watts, 1994). O mesmo autor, usando o exemplo da reparametrização da lei cinética de Arrhenius, argumenta que existe a possibilidade de atribuir um significado físico alternativo aos parâmetros reformulados, e deste modo, evita-se a rejeição do procedimento.

É razoável supor que em situações concretas as formas práticas de transformação dos parâmetros sejam escassas; consequentemente, a definição original (1.3) tem cabimento neste contexto, tendo naturalmente em atenção o facto de o espaço dos parâmetros ser, geralmente, limitado. Uma vantagem desta definição é que se aplica directamente ao caso de resposta multivariada.

Suponhamos então, sem perda de generalidade, que o espaço dos parâmetros é definido por uma região hiperrectangular $\boldsymbol{\theta}^L\leqslant\boldsymbol{\theta}\leqslant\boldsymbol{\theta}^U$. Uma definição possível que decorre da adaptação da de Stromberg e Ruppert é

$$\epsilon_n^+(\boldsymbol{x},f,T,z)=\min_{0\leqslant m\leqslant n}\left(\left\{\frac{m}{n}:b(n,m;T,z)=d(\boldsymbol{\theta}^U,\hat{\boldsymbol{\theta}}),\ \ m\in\{1,\ldots,n\}\right\}\cup\{1\}\right), \quad (1.9)$$

$$\epsilon_n^-(\boldsymbol{x},f,T,z)=\min_{0\leqslant m\leqslant n}\left(\left\{\frac{m}{n}:b(n,m;T,z)=d(\boldsymbol{\theta}^L,\hat{\boldsymbol{\theta}}),\ \ m\in\{1,\ldots,n\}\right\}\cup\{1\}\right), \quad (1.10)$$

e

$$\epsilon_n(\boldsymbol{x},f,T,z)=\min\{\epsilon_n^-(\boldsymbol{x},f,T,z),\epsilon_n^+(\boldsymbol{x},f,T,z)\}. \quad (1.11)$$

Tanto Stromberg e Ruppert (1992) como Sakata e White (1995) apresentam teoremas para a determinação das respectivas versões do ponto de rotura em diversos estimadores robustos.

Por último, refira-se que todos os estimadores possuem ponto de rotura mas não necessariamente função de influência

1.4 O dilema eficiência-robustez

Tal como a teoria clássica, a teoria de estimação robusta pressupõe um modelo paramétrico idealizado para o processo de geração dos erros de medição, mas, em contraste com aquela, não espera que o modelo paramétrico descreva exactamente a realidade. É desejável que o comportamento dos estimadores robustos sob determinado modelo paramétrico assumido seja quase tão bom como o do correspondente estimador óptimo (se existir), isto é, que sejam eficientes. Infelizmente, normalmente os objectivos de eficiência e robustez entram em choque. Assim, é necessário estabelecer um compromisso entre a segurança que se ganha quando numa situação afastada dos pressupostos o estimador não conduz a valores disparatados das estimativas (robustez) e o preço a pagar que se traduz numa perda de eficiência sob a distribuição assumida para os erros. Nisto consiste o dilema eficiência-robustez.

É importante assinalar que este compromisso não tem necessariamente de existir. De facto, Sakata e White (2001, p. 31) mostram que para algumas distribuições do erro a eficiência assimptótica dos estimadores S não-lineares apresenta o comportamento contrário, isto é, aumenta para pontos de rotura superiores.

1.5 Inferência estatística

Qualquer estimativa deve ser acompanhada duma avaliação da sua precisão, nomeadamente através da matriz de covariâncias (o desvio padrão é dado pela raiz quadrada da diagonal principal) dos parâmetros e de intervalos de confiança. Porém, na maioria dos casos não é possível recorrer a expressões teóricas, pois como já referimos, faltam estudos no âmbito da regressão não-linear (existem resultados para os estimadores M e S). Além disso, na presença de *outliers* nos dados, as aproximações assimptóticas não são fiáveis.

Uma via alternativa é o uso da metodologia *bootstrap* introduzida por Efron (1979). (Para uma discussão do *bootstrap* vejam-se Efron e Tibshirani (1993) e Wehrens *et al.* (2000).) O *bootstrap* é um método computacionalmente intensivo dito de reamostragem, pois baseia-se na construção de subamostras da amostra original. Em princípio, é possível estender directamente os procedimentos básicos desenvolvidos para o estimador dos mínimos quadrados a estimadores robustos (Davison e Hinkley, 1997, p. 307). Contudo é necessária cautela. Numa amostra severamente contaminada por *outliers*, mesmo o comportamento de estimadores robustos pode ser mau sob reamostragem, uma vez que é possível obterem-se subamostras com mais de 50% de contaminação, nas quais o estimador ajusta os *outliers* e não os pontos "bons".

Stromberg (1993) efectuou um pequeno estudo de simulação do cálculo da variância dos parâmetros do modelo de Michaelis-Menten obtidos com o estimador MM, tendo observado valores razoáveis em amostras com 20% de contaminação enquanto que em amostras com um grau de contaminação de 40% os valores obtidos encontravam-se inflacionados. Mais recentemente, Stromberg (1997a) provou que o ponto de rotura da estimativa *bootstrap* usual da matriz de covariâncias é $1/n$ independentemente do ponto de rotura do estimador. Quer dizer, basta um ponto aberrante para inutilizar a estimativa *bootstrap*. No entanto, observa que na prática, no caso da proporção de *outliers* ser

baixa, é improvável que a técnica *bootstrap* falhe, pelo que o baixo ponto de rotura não significa necessariamente que a estimativa *bootstrap* básica seja inútil.

A melhoria das propriedades do *bootstrap* é alcançada com o desenvolvimento de variantes adaptadas às características das diferentes classe de estimadores. (Davison e Hinkley, 1997, p. 312) referem modificações para os estimadores M e para o estimador dos desvios absolutos mínimos. Stromberg (1997a) estudou o comportamento por simulação de modificações robustificadas das metodologias de reamostragem *bootstrap* e *jackknife* tradicionais para estimativa da matriz de covariâncias no contexto de localização. Uma variante *bootstrap* e outra *jackknife* distinguem-se, acomodando até $30\% - 40\%$ de *outliers*. Finalmente, sublinhe-se que a avaliação do potencial das alternativas referidas no âmbito da regressão não-linear carece de ser investigada.

1.6 Descrição dos capítulos

O capítulo 2 é devotado à descrição dos estimadores robustos quer no caso univariado quer no caso multivariado, expõe os escassos algoritmos desenvolvidos especificamente para o modelo de regressão não-linear, assim como o contexto que conduziram à escolha de uma variante do método de evolução diferencial. Por fim, descreve alguns aspectos da implementação computacional dos diversos estimadores em competição.

O capítulo 3 apresenta a metodologia das simulações de Monte Carlo.

O capítulo 4 descreve os resultados e discussão do caso univariado. Segue-se-lhe o capítulo 5 dedicado ao caso multivariado.

Finalmente, o capítulo 6 fecha o presente trabalho com conclusões e o trabalho de investigação futuro a desenvolver.

Capítulo 2

Apresentação das classes de estimadores

Para uma melhor compreensão das famílias de estimadores consideradas nas experimentações de Monte Carlo que irão ser explicados posteriormente começar-se-á por rever um conjunto de critérios de qualidade dos estimadores. Uma vez que quer as simulações quer os algoritmos usados para o cálculo numérico dos estimadores são de natureza computacionalmente intensiva, o factor limitante da abrangência dos estudos de caso é a capacidade de cálculo disponível. Por isso, a selecção é necessariamente não exaustiva.

2.1 Critérios de avaliação da qualidade dos estimadores

Nesta secção far-se-á uma revisão dos principais critérios utilizados para a avaliação da qualidade de determinado estimador.

Não-enviesamento

Seja T um estimador para o parâmetro θ. Chama-se enviesamento de T à diferença entre o valor esperado do estimador, $E(T)$, e o verdadeiro valor do parâmetro, θ, ou seja

$$\text{Enviesamento} = E(T) - \theta.$$

Se o seu enviesamento for nulo, ou seja, $E(T) = \theta$, um estimador diz-se *não-enviesado ou centrado*; caso contrário, diz-se então *enviesado*. Segundo Murteira *et al.* (2001, p. 345) a ideia é que "... qualquer que seja a dimensão da amostra, um 'bom' estimador deve fornecer, 'em média', estimativas exactas, isto é, coincidentes com o verdadeiro valor do parâmetro".

Eficiência

Para dois estimadores não-enviesados T_1 e T_2 de um parâmetro θ com variâncias σ_1^2 e σ_2^2, respectivamente, diz-se que o estimador T_1 é *relativamente mais eficiente* do que T_2 quando $\sigma_1^2 \leqslant \sigma_2^2$.

Um modo bastante utilizado para medir a eficiência de T_2 relativamente a T_1 passa pelo cálculo do quociente

$$\frac{\sigma_1^2}{\sigma_2^2}.$$

Como a variância de um estimador mede a dispersão em torno do seu valor esperado, não pode utilizar-se como critério para comparar estimadores enviesados. Uma forma possível de resolver o problema da extensão do conceito de eficiência a estimadores enviesados consiste em usar o critério do erro quadrático médio (MSE) de T, $E[(T-\theta)^2]$.

Consistência

Um estimador T do parâmetro θ diz-se *consistente* quando, para qualquer número real $\delta > 0$,

$$\lim_{n\to\infty} \text{Probabilidade}(|T - \theta| < \delta) = 1,$$

onde n representa a dimensão da amostra.

Esta definição significa que a precisão de um estimador consistente é tanto maior quanto maior for a dimensão da amostra. Observe-se ainda que, ao invés dos critérios anteriormente descritos, a consistência é uma propriedade assimptótica.[1]

Realce-se que a importância deste critério "... advém sobretudo da ideia de que um estimador que não seja consistente é um estimador que não deve ser utilizado" (Murteira *et al.*, 2001, p. 353). Importa não esquecer, no entanto, que esta perspectiva é duramente criticada por Le Cam e Yang como se viu na página 2.

2.2 Modelos com resposta univariada

2.2.1 Estimador dos mínimos quadrados

Consideremos o modelo (1.1) reescrito na forma

$$\epsilon_i = y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta}), \qquad i = 1, 2, \ldots, n, \tag{2.1}$$

e seja $\hat{\boldsymbol{\theta}}$ o vector das estimativas dos parâmetros. O estimador dos LS minimiza a soma do quadrado dos resíduos, ou seja,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \epsilon_i^2(\boldsymbol{\theta}), \tag{2.2}$$

e coincide com estimador de máxima verosimilhança no pressuposto dos erros ϵ_i serem independentes e identicamente distribuídos e seguirem a distribuição Gaussiana com média nula e variância desconhecida σ_ϵ^2. Uma estimativa centrada do desvio padrão do erro de medição é dada por

$$s = \left(\frac{1}{n - n_p} \sum_{i=1}^{n} \epsilon_i^2(\hat{\boldsymbol{\theta}}) \right)^{1/2}. \tag{2.3}$$

Como cada erro contribui de forma quadrática para a função objectivo, a presença de *outliers* nas observações — que se traduzem erros de elevado valor absoluto — tem um efeito severo nas estimativas obtidas a partir deste estimador.

A forma da respectiva função de influência (tomando ϵ como variável independente) é a função identidade. Sem qualquer surpresa, note-se que esta não limita de forma alguma o efeito de um ponto anómalo nas estimativas dos parâmetros, isto é, $\epsilon \to \infty \Rightarrow \text{IF} \to \infty$. Por outro lado, o ponto de rotura é $1/n$, quer para a definição de Stromberg e Ruppert

[1] O resultado da determinação da eficiência relativa no limite, quando a dimensão da amostra aumenta indefinidamente, designa-se por *eficiência relativa assimptótica.*

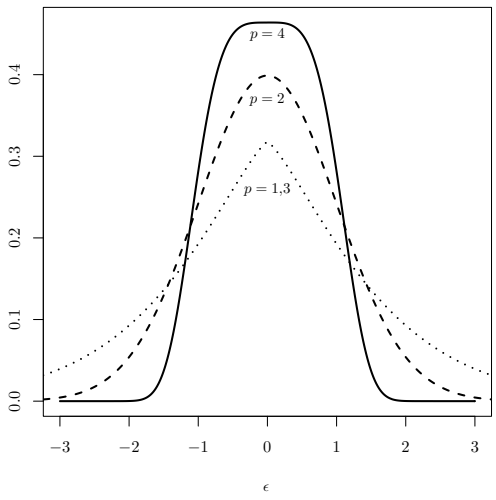


Figura 2.1 Função densidade da distribuição exponencial-potência para diferentes valores de p.

quer para a de Sakata e White. Ou seja, basta a presença de um *outlier* na amostra para que o estimador LS "rompa", embora tal não tenha necessariamente de ocorrer.

Por isso, supondo razoável tomar como linha de orientação o estudo na área da química analítica referido em Phillips e Eyring (1983) devido a Clancy (1947), no qual menos de 15% de 250 distribuições baseadas em 50 000 análises podem ser consideradas Gaussianas — os desvios são atribuíveis à ocorrência de *outliers* ou a características intrínsecas da distribuição dos erros —, e atendendo a que o estimador LS é susceptível ao efeito de *dissimulação* (*masking*) — em que o efeito de múltiplos *outliers* implica que estes não sejam detectados pelos procedimentos de diagnóstico usuais —, o uso exclusivo do estimador LS envolve um perigo considerável.

2.2.2 Estimadores de norma L_p

Quando se assume que a lei dos erros de medição segue a distribuição exponencial-potência, o estimador de máxima verosimilhança (ML) resultante designa-se por estimador de norma L_p. A função densidade de probabilidade desta distribuição é dada por

$$g(\epsilon_i;\mu,\phi,p) = \frac{1}{\phi\Gamma(1+1/p)2^{1+1/p}} \exp\left(-\frac{1}{2}\left|\frac{\epsilon_i-\mu}{\phi}\right|^p\right), \quad \text{com} \quad \phi>0, \quad 1\leqslant p<\infty, \tag{2.4}$$

onde Γ é a função Gama, μ é o parâmetro de localização que corresponde à média, ϕ é o parâmetro de escala, e p é o parâmetro de forma, o qual determina a *kurtosis* da distribuição: para $p<2$ a distribuição é *leptokurtica* (zona central mais "pontiaguda" com caudas mais "espessas" que a distribuição Gaussiana), e para $p>2$ a distribuição é *platikurtica* (caudas mais "finas" com zona central mais "achatada" do que a distribuição Gaussiana). A figura 2.1 ilustra a forma da função densidade nestas situações. Assinale-se que a distribuição Gaussiana e a distribuição de Laplace correspondem a casos particulares da distribuição exponencial-potência, quando $p=2$ e $p=1$, respectivamente.

Assumindo que os erros de medição são independentes, de média nula e com variância

constante, a *função de verosimilhança* (condicionada por $\boldsymbol{y}$) é definida por

$$L(\boldsymbol{\theta}, \phi, p; \boldsymbol{y}) = [\phi\Gamma(1+1/p)2^{1+1/p}]^{-n} \exp\left(-\frac{1}{2\phi}\sum_{i=1}^{n}|y_i - f(x_i, \boldsymbol{\theta})|^p\right)$$

onde a ordem dos símbolos entre parêntesis em L enfatizam o facto de os parâmetros serem condicionados pelas observações. Na generalidade das aplicações é mais fácil trabalhar com o logaritmo da função de verosimilhança, que designaremos por l,

$$l = \ln L = -n\ln[\phi\Gamma(1+1/p)2^{1+1/p}] - \frac{1}{2\phi}\sum_{i=1}^{n}|y_i - f(x_i, \boldsymbol{\theta})|^p. \tag{2.5}$$

Maximizando l com respeito aos parâmetros, obtêm-se as estimativas de $\boldsymbol{\theta}$, ϕ, e p que maximizam a probabilidade de ocorrer os valores específicos de $\boldsymbol{y}$ observados na amostra.

Gonin e Money (1989, p. 222) referem que este procedimento "embora teoricamente correcto não é muito praticável", não especificando porquê. Assim, é prática corrente estabelecer à partida um valor particular para p. Logo, neste caso, a maximização de l com respeito a $\boldsymbol{\theta}$ conduz a

$$\hat{\boldsymbol{\theta}} = \arg\min \sum_{i=1}^{n}|y_i - f(x_i, \boldsymbol{\theta})|^p, \tag{2.6}$$

que pode ser reescrita como

$$\hat{\boldsymbol{\theta}} = \arg\min \sum_{i=1}^{n}\rho(\epsilon_i), \quad \text{onde} \quad \rho(u) = |u|^p.$$

Note-se que o estimador de norma L_p inclui como casos particulares o estimador LS para $p = 2$, e o estimador dos LAD para $p = 1$; de igual modo, a distribuição exponencial-potência coincide, respectivamente, com a distribuição Gaussiana e com a distribuição de Laplace.

Para $1 \leqslant p < 2$ este estimador é resistente a *outliers* no espaço da variável dependente (Gonin e Money, 1989, p. 11). Contudo apresenta um baixo ponto de rotura porque é sensível à ocorrência de *outliers* no espaço das variáveis independentes: sob as condições dos teoremas 3.6(a) e 3.6(b) apresentados no artigo de Sakata e White, o ponto de rotura é $1/n$, idêntico ao do estimador LS.

A função de influência é proporcional a $\psi = \rho'(u)$[2] (Huber, 1996, p. 14), ou seja,

$$\text{IF} \propto \begin{cases} \operatorname{sgn}\epsilon & \text{se } p = 1 \\ p\,\epsilon|\epsilon|^{p-2} & \text{se } p > 1, \end{cases}$$

onde a função sinal sgn toma os valores 1, 0, ou -1, consoante o argumento é positivo, zero ou negativo. A figura 2.2 na próxima página mostra a função de influência dos estimadores LAD, LS, e de norma L_p para $p = 1{,}5$. Verificamos que a função de influência do estimador LAD é limitada em todo o seu domínio, o que se traduz igualmente no

[2]Na literatura estatística a função ρ' é comummente designada por ψ.

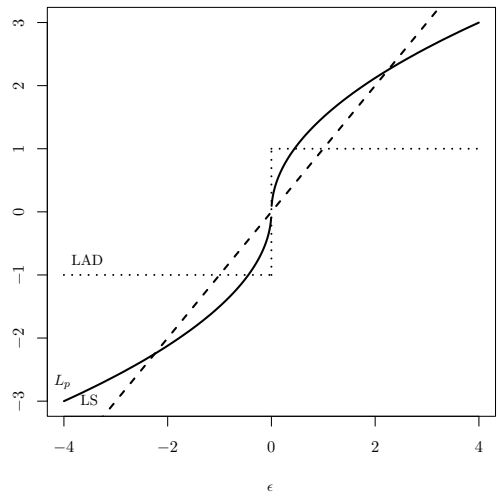


Figura 2.2 Funções ψ dos estimadores LAD, LS, e de norma L_p ($p = 1{,}5$).

impacto limitado de um ponto arbitrário nas estimativas. Por outro lado, embora não limitada, a função de influência do estimador de norma L_p ($p = 1{,}5$) é inferior em valor absoluto à do estimador LS se o valor de $|\epsilon|$ exceder aproximadamente 2, o que determina uma sensibilidade inferior à presença de um ponto anómalo.

Como notámos anteriormente, a abordagem usual na literatura especifica à partida um valor para p. Coloca-se agora a questão da escolha desse valor. Gonin e Money (1989, pp. 226–227) apresentam um procedimento adaptativo heurístico de selecção, o qual, em cada iteração k, mediante a aplicação de certas regras baseadas na curtose dos resíduos obtidos pela estimação com um valor particular de p^k, gera uma nova estimativa p^{k+1}.

No contexto Bayesiano Militký e Čáp (1987) apresentam uma abordagem que estima *simultaneamente* os parâmetros do modelo e p. O procedimento de optimização sugerido separa o cálculo das estimativas de p e $\boldsymbol{\theta}$ em dois ciclos: no ciclo exterior decorre a estimação de p, enquanto no ciclo interior são determinados os valores das estimativas (2.6) correspondentes ao valor tomado por p no ciclo interior.

Neste trabalho optou-se pela abordagem análoga no contexto frequencista, a qual minimiza a função log-verosimilhança (2.5). Na nossa experiência, a optimização directa desta expressão é exequível com os algoritmos estocásticos referidos na secção 2.4 na página 30.

A aplicação do método da verosimilhança concentrada (Seber e Wild, 1989, pp. 37–42) possibilita a eliminação do parâmetro ϕ do procedimento de estimação. Assim, para $\boldsymbol{\theta}$ e p fixos a expressão (2.5) é maximizada quando

$$\frac{\mathrm{d}l}{\mathrm{d}\phi} = -\frac{n}{\phi} + \frac{p}{2\phi^{p+1}} \sum_{i=1}^{n} |\epsilon_i|^p = 0,$$

e, portanto,

$$\phi^p = \frac{p}{2n} \sum_{i=1}^{n} |\epsilon_i|^p.$$

Substituindo a equação anterior na equação (2.5), e procedendo a simplificações, pode-

mos verificar que a função de verosimilhança concentrada é

$$l = -n\left\{\ln\left[\left(\frac{p}{2n}\right)^{1/p}\Gamma\left(1+\frac{1}{p}\right)2^{1+1/p}\left(\sum_{i=1}^{n}|\epsilon_i|^p\right)^{1/p}\right]+\frac{1}{p}\right\}.$$

Então,

$$(\hat{\boldsymbol{\theta}}^T,\hat{p})^T = \arg\max_{\boldsymbol{\theta},p} l = \arg\min_{\boldsymbol{\theta},p}\left\{\ln\left[\left(\frac{p}{2n}\right)^{1/p}\Gamma\left(1+\frac{1}{p}\right)2^{1+1/p}\left(\sum_{i=1}^{n}|\epsilon_i|^p\right)^{1/p}\right]+\frac{1}{p}\right\} \tag{2.7}$$

e

$$\hat{\phi} = \left(\frac{\hat{p}}{2n}\sum_{i=1}^{n}|\hat{\epsilon}_i|^{\hat{p}}\right)^{1/\hat{p}}, \quad \text{com} \quad \hat{\epsilon}_i = y_i - f(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}). \tag{2.8}$$

Dadas as propriedades dos estimadores ML, a estimativa da variância do erro de medição, $\hat{\sigma}^2_\epsilon$, obtém-se substituindo ϕ pelo valor da respectiva estimativa na expressão (Box e Tiao, 1973, p.)

$$\sigma^2_\epsilon = 2^{2/p}\frac{\Gamma(3/p)}{\Gamma(1/p)}\phi^2. \tag{2.9}$$

Finalmente, refira-se que como a variante robusta destes estimadores restringe p ao intervalo [1, 2[, o seu cálculo implica o uso de um algoritmo de optimização que restrinja o domínio de procura das variáveis.

2.2.3 Estimadores M

Apesar de não serem considerados nas simulações com resposta univariada deste estudo, os estimadores M entram na definição dos estimadores MM e τ, optando-se por esse motivo pela sua descrição a par das outras famílias desta secção.

Introduzidos por Huber (1973) no contexto de regressão, a ideia básica consiste em substituir o quadrado dos resíduos em (2.2) por outra função construída de forma a que o estimador associado seja tão robusto e eficiente quanto possível. Matematicamente

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}}\sum_{i=1}^{n}\rho(\epsilon_i(\boldsymbol{\theta})), \tag{2.10}$$

onde ρ é uma função simétrica com um mínimo em zero.

Assumamos que as variáveis explicativas $\boldsymbol{x}$ constituem uma amostra aleatória de uma distribuição multivariada K, e são independentes do erro de medição ϵ com distribuição F. Suponhamos, agora, que K é simétrica e $\psi = \rho'$ ímpar. No caso da regressão linear, a função de influência pode ser escrita da seguinte forma:

$$\frac{\psi(\epsilon)}{E_F[\psi'(\epsilon)]}\{(E_K[\boldsymbol{x}\boldsymbol{x}^T])^{-1}\boldsymbol{x}\}, \tag{2.11}$$

onde E designa a esperança matemática. Note que a expressão acima é um produto de dois termos: um escalar que descreve o efeito induzido pelos resíduos proporcional a ψ, e um vector função de $\boldsymbol{x}$ que reflecte o efeito de um ponto arbitrário.

A razão do nome deve-se ao facto destes estimadores estarem relacionados com os estimadores ML. Estes últimos são aproximadamente estimadores centrados de variância mínima. A sua falta de robustez deve-se a que a respectiva função de influência é tipicamente ilimitada. Assim, em termos muito gerais, uma abordagem possível para alcançar os objectivos de eficiência e robustez consiste em tomar ψ proporcional à derivada da função log-verosimilhança definida pela função densidade g da distribuição particular assumida para as observações, isto é,

$$\psi(y) = -g'(y)/g(y),$$

modificando-a de modo a torná-la contínua e limitada (Staudte e Sheather, 1990, p. 112).

A literatura da estatística robusta apresenta inúmeras funções ψ obtidas a partir de diferentes considerações de carácter assimptótico. Limitamos esta discussão às funções de Huber (Huber, 1964)

$$\rho(u) = \begin{cases} \frac{u^2}{2} & \text{se } |u| \leqslant c \\ c\left(|u| - \frac{c}{2}\right) & \text{se } |u| > c \end{cases} \qquad \text{para a qual} \qquad \psi(u) = \begin{cases} -c & \text{se } u < -c \\ u & \text{se } |u| \leqslant c \\ c & \text{se } u > c, \end{cases}$$

e à função redescendente[3] de Hampel (Hampel, 1974)

$$\rho(u) = \begin{cases} \frac{u^2}{2} & \text{se } |u| < a \\ a\left(|u| - \frac{a}{2}\right) & \text{se } a \leqslant |u| < b \\ ab - \frac{a^2}{2} + (c-b)\frac{a}{2}\left[1 - \left(\frac{c-|u|}{c-b}\right)^2\right] & \text{se } b \leqslant |u| \leqslant c \\ ab - \frac{a^2}{2} + (c-b)\frac{a}{2} & \text{se } |u| > c \end{cases}$$

para a qual

$$\psi(u) = \begin{cases} u & \text{se } |u| < a \\ a \operatorname{sgn} u & \text{se } a \leqslant |u| < b \\ a \operatorname{sgn} u \dfrac{c-|u|}{c-b} & \text{se } b \leqslant |u| \leqslant c \\ 0 & \text{se } |u| > c. \end{cases}$$

Designaremos os pontos a, b, e c por nós. A figura 2.3 na próxima página representa as expressões anteriores.

À luz das considerações anteriores, pode ver-se que ambas as funções possuem uma porção central idêntica à do estimador LS responsável pela eficiência destes estimadores sob a distribuição normal. A função de Huber ao limitar a respectiva função de influência para $|u| > c$ diminui o peso dos resíduos mais elevados comparativamente ao estimador LS resistindo assim ao efeito de *outliers*. Por outro lado, além disso, a função

[3] A designação redescendente aplica-se a funções ρ cuja respectiva função de influência apresenta um comportamento decrescente longe da origem. (Tipicamente, a função de influência dos estimadores redescendentes anula-se a partir de um valor crítico de $|u|$.)

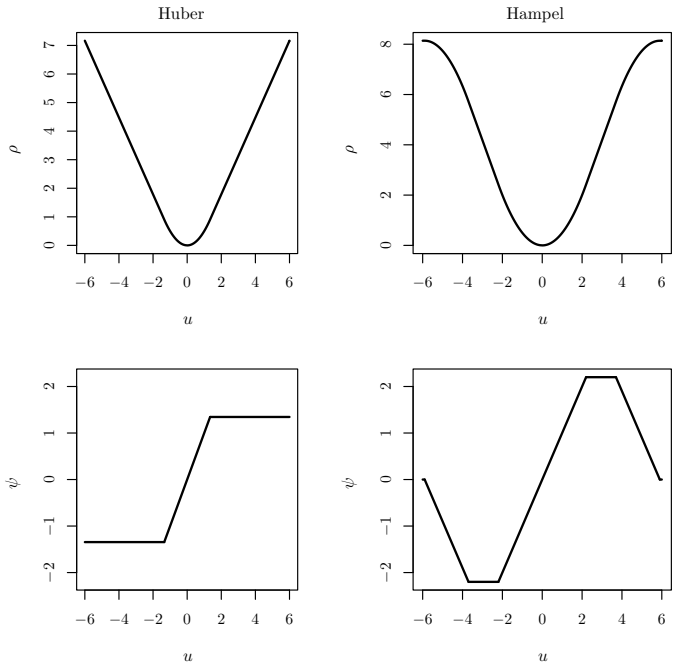


Figura 2.3 Funções ρ e ψ dos estimadores M de Huber ($c = 1{,}345$) e Hampel ($a = 2{,}2$, $b = 3{,}7$, $c = 5{,}9$).

de Hampel rejeita completamente os *outliers* a partir de $|u| > c$. O decréscimo suave para 0 — em contraste com uma transição abrupta — da função de influência define uma zona de transição que permite considerar a informação presente em *outliers* moderados, melhorando-se assim a eficiência do estimador.

Os valores numéricos dos nós determinam o equilíbrio entre os objectivos conflituosos de eficiência — em geral, sob a distribuição Gaussiana — e robustez, sendo usualmente escolhidos com base em estudos de carácter assimptótico. Por exemplo, o aumento do valor de c no estimador M de Huber melhora a sua eficiência, mas torna-o sensível a uma gama de *outliers* moderados mais alargada, o que diminui a sua robustez. Note-se que quando $c \to \infty$, o estimador M de Huber tende para o estimador LS.

É importante enfatizar que a função de influência não é limitada na direcção de $\boldsymbol{x}$, ao contrário do que ocorre na direcção dos resíduos. Deste modo, as estimativas mostram grande sensibilidade à presença de observações que sejam *outliers* no espaço das variáveis explicativas.

Tendo em atenção as expressões de ρ e ψ verificamos que apresentam algumas "patologias" que colocam dificuldades aos algoritmos determinísticos de optimização usuais na resolução de (2.10), nomeadamente:

- a segunda derivada de ambas as funções é descontínua[4] nos pontos a, b, e c —

[4]Existem excepções com 1.ª e 2.ª derivadas contínuas como a função de Fair,

$$\rho(u) = c^2(|u|/c - \ln(1 + |u|/c)) \quad \text{com} \quad c > 0,$$

ou a função de Beaton e Tukey (1974),

$$\rho(u) = \begin{cases} \frac{1}{3}a^2\{1 - [1 - (u/a)^2]^3\} & \text{se } |u| \leqslant a \\ \frac{1}{3}a^2 & \text{se } |u| > a. \end{cases}$$

existindo inclusivamente funções com derivada de primeira ordem descontínua; e

- a função ρ de Hampel, *não* é convexa, o que, independentemente das características do modelo, implica a existência de múltiplos óptimos locais. Esta característica é comum a todas as funções redescendentes.

Por outro lado, as estimativas dependem das unidades de medida das observações, quer dizer, os estimadores M assim definidos não são equivariantes à escala de y. Para isso é necessário normalizar ϵ_i por um factor de escala σ. Logo, a partir da expressão (2.10) podemos escrever

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho\left(\frac{\epsilon_i(\boldsymbol{\theta})}{\sigma}\right). \tag{2.12}$$

Na prática, raramente o valor de σ é conhecido à partida. Por isso, uma prática corrente consiste em estimá-lo a partir duma medida robusta de dispersão aplicada aos resíduos da iteração corrente ou anterior do algoritmo de optimização (Seber e Wild, 1989, p. 651; Venables e Ripley, 1999, p. 169), tal como o estimador mediana dos desvios absolutos em relação à mediana (MAD)

$$\hat{\sigma} = 1{,}4826 \operatorname*{med}_i |\hat{\epsilon}_i - \operatorname*{med}_j \hat{\epsilon}_j|.$$

Uma abordagem alternativa designada por proposta 2 de Huber consiste na estimação simultânea de σ e $\boldsymbol{\theta}$ pela minimização duma expressão da forma (Huber, 1996, pp. 37 e 38; Dutter e Huber, 1981, p. 80)

$$\sum_{i=1}^{n} \rho\left(\frac{\epsilon_i(\boldsymbol{\theta})}{\sigma}\right)\sigma + a\sigma, \qquad \sigma > 0, \tag{2.13}$$

onde $a = (n - n_p)E_\Phi[u\psi(u) - \rho(u)]$ é uma constante que assegura a consistência da estimativa de σ sob erros Gaussianos, e ρ é uma função convexa *(o que exclui as funções ρ redescendentes)* que satisfaz as seguintes condições: $\rho(u) \geqslant 0$, $\rho(0) = 0$, $\rho(u)/u$ é convexa para $u < 0$ e côncava para $u > 0$, e é duplamente diferenciável. Por último, Lawrence e Arthur (1990) tomam para o valor de σ o desvio padrão dos resíduos de um ajuste preliminar obtido pelo método dos mínimos quadrados. Midi (1999) propõe uma versão robusta deste procedimento, na qual o estimador LS é substituído pelo estimador mais robusto LAD, e $\hat{\sigma}$ é definido por $1{,}5 \operatorname{med}_i|\hat{\epsilon}_i|$.

Olhemos agora para algumas propriedades do caso da regressão não-linear. Neugebauer (1996, p. 23) mostrou que a diferença da função de influência relativamente ao caso linear está no facto de o termo entre chavetas na expressão (2.11) ser função do gradiente $\partial f(\boldsymbol{x}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$. (Note-se que no modelo linear $\partial f(\boldsymbol{x}, \boldsymbol{\theta})/\partial\boldsymbol{\theta} = \boldsymbol{x}$.) Portanto, uma influência limitada depende não só de uma função ψ limitada mas também da forma do modelo e suas derivadas. Assim, é possível no contexto não-linear que os estimadores M não sejam perturbados por *outliers* em $\boldsymbol{x}$, ou que exibam sensibilidade a pontos situados no interior do globo de $\boldsymbol{x}$ nos dados. Por outras palavras, um ponto de influência[5] não é

[5] O conceito de influência mede a sensibilidade das estimativas dos parâmetros do modelo em relação a perturbações na variável dependente, em uma qualquer observação da amostra. Chama-se *ponto de influência* a uma observação com influência muito acentuada, caso exista, relativamente à grande maioria das restantes observações da amostra.

necessariamente um *outlier* em $\boldsymbol{x}$. O mesmo autor apresenta vários exemplos ilustrativos no capítulo 5. Repare-se que estas considerações se aplicam igualmente aos estimadores LS, LAD, e de norma L_p, visto que constituem instâncias dos estimadores M.

Finalmente, os estimadores M partilham com o estimador de norma L_p a falta de robustez a grupos de *outliers* no espaço das variáveis independentes. Sakata e White (1995) mostram nos teoremas referidos na secção anterior que o ponto de rotura, tal como para o estimador de norma L_p, é igualmente $1/n$ e chamam a atenção para o perigo do uso destes estimadores. Contudo, alguns investigadores argumentam que não é clara a presença de *outliers* nas variáveis independentes na maioria das situações experimentais, — em que os valores de $\boldsymbol{x}$ não são observados mas fixos pelo experimentalista —, pelo que é defensável a sua utilização neste contexto. Porém, Croux *et al.* (1994) identificaram *outliers* em $\boldsymbol{x}$ num caso real em que se usou planeamento de experiências, o que conduz os autores a observar que um procedimento de planeamento de experiências pode igualmente conduzir a pontos com um afastamento grande (no espaço das variáveis independentes) em relação aos outros pontos da amostra.

2.2.4 Estimadores de elevado ponto de rotura: mínimos quadrados aparados e mediana mínima dos quadrados

De modo a alcançar robustez à ocorrência de *múltiplos outliers* no espaço das variáveis independentes e no espaço da variável dependente, Rousseeuw (1984) propôs o estimador da LMS, dado por

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \operatorname*{med}_i \epsilon_i^2(\boldsymbol{\theta}). \tag{2.14}$$

Stromberg (1995) apresenta a prova de que este estimador é fracamente consistente no âmbito da regressão não-linear.

Uma limitação deste estimador prende-se com a sua baixa eficiência sob a distribuição Gaussiana. No contexto da regressão linear apresenta uma eficiência assimptótica de 0%. Uma vez que não existem na literatura resultados para regressão não-linear, admite-se que neste enquadramento os resultados obtidos para regressão linear são semelhantes.

Para ultrapassar este problema, o mesmo autor desenvolveu o estimador dos LTS, definido como,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{h} \{\epsilon^2(\boldsymbol{\theta})\}_{i:n} \tag{2.15}$$

onde $\{\epsilon^2\}_{1:n} \leqslant \ldots \leqslant \{\epsilon^2\}_{n:n}$ são os resíduos quadrados ordenados, $\{\,\cdot\,\}_{i:n}$ representa a i-ésima estatística de ordem duma colecção de n números, e h é o número pré-fixado de observações não aparadas da amostra, escolhido de forma conveniente, e designa-se por cobertura. Define-se a fracção de aparamento, α, tal que $h = n(1-\alpha)$, onde h é arredondado ao inteiro mais próximo. De notar que a única diferença entre este estimador e o estimador LS reside no facto de na definição 2.15 a soma excluir os resíduos de valor absoluto mais elevado, tornando assim o estimador resistente a *outliers*. Chen *et al.* (1997) mostraram que o estimador LTS não-linear é fortemente consistente.

Um elevado ponto de rotura é alcançado tomando $h \approx n/2$. Infelizmente, a eficiência assimptótica deste estimador sob a distribuição Gaussiana, igualmente no contexto da regressão linear, é apenas 7%.

O objectivo de que um estimador robusto deve possuir um desempenho semelhante do estimador LS sob a distribuição Gaussiana em condições livres de contaminação, implícito nas considerações anteriores, é questionável. Historicamente, o uso generalizado do estimador LS deve-se a ser especialmente tratável matematicamente, e decorre daqui a ênfase colocada na distribuição Gaussiana — que é o mecanismo de geração de dados para o qual o estimador LS é óptimo — na literatura. Existe, contudo, evidência de que estas condições raramente são satisfeitas na prática (Rousseeuw e Leroy, 1987, p. 41;). Por outro lado, se a utilização destes estimadores for feita num quadro de análise de dados — um subproduto dos estimadores robustos é poderem ser utilizados como ferramenta de diagnóstico — a questão da eficiência estatística não é crucial (Rousseeuw e Leroy, 1987, p. 188).

Em regressão linear, o estimador LMS e o estimador LTS com $h \approx n/2$ têm um ponto de rotura de 0,5 — o valor máximo teórico alcançável. Em contraste com esta situação, em regressão não-linear é possível um valor inferior (Stromberg e Ruppert, 1992; Sakata e White, 1995). Por outro lado, o valor máximo teórico excede 0,5 em $1/n$, o que é notável tendo em atenção que configura uma situação em que o número de pontos "maus" na amostra excede o número de pontos "bons". No modelo de Michaelis-Menten, Stromberg e Ruppert (1992) explicam este efeito, observando que o modelo é constrangido a passar pelo ponto $(0, 0)$, o que gera um ponto "bom" adicional.

Tem sido argumentado no âmbito da regressão linear que a generalidade dos estimadores de elevado ponto de rotura, por exemplo 0,5, podem ser influenciado por certas configurações de *outliers*, que conduzem a um desempenho pobre, ainda que a fracção de contaminação seja consideravelmente inferior ao ponto de rotura (Ryan, 1997, pp. 358–360). A ideia básica do exemplo apresentado é a existência de um segundo hiperplano — constituído pelo alinhamento dos *outliers* com pontos "bons" — que ajusta quase exactamente metade dos dados. Nesta situação, o uso de um ponto de rotura de 0,5 conduz à equação de regressão deste hiperplano, desviando o estimador da equação apropriada. De modo a ultrapassar a deficiência que este comportamento confere ao uso isolado de um ponto de rotura elevado, Ryan advoga o uso de múltiplos pontos de rotura num estimador de elevado ponto de rotura e comparar os resultados.

Em contraste com este ponto de vista, Yohai e Zamar (1988) argumentam que a possível existência de mais que uma estrutura linear muito provavelmente "reflectirá alguma estrutura intrínseca e com significado dos dados; por isso,a sua identificação melhora a análise e entendimento nos dados". Assim, constitui uma característica desejável e não uma fraqueza dos estimadores de elevado ponto de rotura. Por outro lado, Yohai e Zamar referem a possibilidade de utilizar um ponto de rotura arbitrário inferior a 0,5.

No caso do estimador LTS este procedimento implica a utilização de diferentes percentagens de aparamento. Coloca-se agora a questão de determinar qual o aparamento a utilizar. O caso ideal seria aparar o número exacto de *outliers* conservando a totalidade dos pontos "bons" da amostra. O uso dum aparamento inferior corresponde a uma situação em que se aplica o estimador LS a uma amostra que ainda incluí *outliers*, se bem que em número inferior, enquanto que um aparamento em excesso do número de *outliers* pode conduzir ao comportamento referido anteriormente. Ryan (1997, pp. 372–377) sugere um procedimento *ad hoc* no qual o estimador LTS é usado sequencialmente com valores de α sucessivamente superiores. O processo é iniciado com o estimador LS, e termina quando a redução da função objectivo para dois valores de α consecutivos,

formulada como

$$n\,\frac{\sum_{i=1}^{n(1-\alpha^{(k)})}\epsilon_i^2(\hat{\boldsymbol{\theta}}_{\mathrm{LTS}}^{(k)})-\sum_{i=1}^{n(1-\alpha^{(k+1)})}\epsilon_i^2(\hat{\boldsymbol{\theta}}_{\mathrm{LTS}}^{(k+1)})}{\sum_{i=1}^{n}\epsilon_i^2(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}^{(0)})},$$

alcança um valor crítico — o autor sugere 3.

2.2.5 Estimador MM

Yohai (1987) introduziu o estimador MM que combina elevado ponto de rotura com elevada eficiência quando os erros seguem a distribuição Gaussiana. A ideia por detrás do estimador MM é iniciar o procedimento de cálculo iterativo da estimativa dum estimador eficiente sob a distribuição Gaussiana com uma estimativa robusta com elevado ponto de rotura, sob condições que possibilitam ao estimador eficiente herdar o ponto de rotura do estimador de elevado ponto de rotura.

O estimador MM define-se em três passos:

1. Em primeiro lugar calcula-se uma estimativa consistente $\hat{\theta}_0$ com elevado ponto de rotura, digamos 0,5. Isto pode ser conseguido usando o estimador LMS ou o estimador LTS com percentagem de aparamento de 50%, atendendo a que ambos são consistente no âmbito da regressão não-linear. Stromberg (1993) sugere a estimativa LMS, porém, mais recentemente o mesmo autor refere a existência de alguma evidência de que a estimativa LTS pode ser mais apropriada (Stromberg, 1997b).

2. Em seguida, determinam-se os resíduos

$$\epsilon_i(\hat{\boldsymbol{\theta}}_0)=y_i-f(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_0),\qquad i=1,2,\ldots,n, \tag{2.16}$$

   e calcula-se a estimativa M de escala, s_n, dada pela solução da equação algébrica não-linear

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{\epsilon_i(\hat{\theta}_0)}{s_n}\right)=b,\qquad \text{com}\quad \rho_0(u)=\rho(u/k_0), \tag{2.17}$$

   onde k_0 e b são constantes e b é tal que $b/a=0{,}5$ com $a=\max\rho_0(u)$. A função ρ deve satisfazer as seguintes condições:

   a) é simétrica, continuamente diferenciável, e $\rho(0)=0$; e
   b) existe um $c>0$ tal que ρ é estritamente crescente em $[0,c]$ e constante em $[c,\infty)$.

   Estas condições implicam que ψ tem de ser redescendente.

3. Finalmente, calcula-se uma estimativa M $\hat{\boldsymbol{\theta}}_1$ dada por *qualquer mínimo local* de

$$S(\boldsymbol{\theta})=\sum_{i=1}^{n}\rho_1\left(\frac{\epsilon_i(\boldsymbol{\theta})}{s_n}\right), \tag{2.18}$$

   que satisfaça $S(\hat{\boldsymbol{\theta}}_1)\leqslant S(\hat{\boldsymbol{\theta}}_0)$. A forma óbvia de assegurar esta condição consiste em inicializar o algoritmo de cálculo com a estimativa $\hat{\boldsymbol{\theta}}_0$ determinada no 1º passo.

Tabela 2.1 Valores de k_1 para várias eficiências

Eficiência (%)	80	85	90	95	96	97	98	99
k_1	0,4950	0,5704	0,6877	0,9014	0,9687	1,0524	1,1642	1,3402

Fonte: Stromberg (1993, p. 238).

O factor de escala s_n é o calculado no passo anterior, e $\rho_1(u) = \rho(u/k_1)$, onde k_1 é uma constante que determina a eficiência assimptótica da estimativa sob erros Gaussianos.

Stromberg (1993) modifica este passo ao calcular uma estimativa M adicional que usa como ponto de partida a estimativa LS dos dados. A estimativa que apresentar o menor valor da função objectivo constitui a estimativa MM final.

Stromberg (1993) usa a função ρ de Hampel com $(a, b, c) = (1{,}5, 3{,}5, 8)$ e $k_0 = 0{,}212$. A tabela 2.1 apresenta os valores de k_1 determinados por Stromberg correspondentes às eficiências especificadas. Uma vez que a estimativa M é um mínimo local e se dispõe de um ponto de partida de boa qualidade, pode ser empregue qualquer algoritmo de optimização determinístico clássico. De uma maneira análoga a Edgar *et al.* (2001, p. 385), neste caso assume-se implicitamente o pressuposto que os pontos de descontinuidade da 2ª derivada apresentados pela generalidade das funções ρ não são percorridos no decorrer do procedimento de optimização.

Stromberg (1993) observa que é razoável aproximar a matriz das covariâncias dos parâmetros assimptótica pela expressão

$$\frac{\frac{1}{n-p}\hat{s}_n^2 \sum_{i=1}^n \left[\psi_1\left(\frac{\epsilon_i(\hat{\boldsymbol{\theta}}_{\mathrm{MM}})}{\hat{s}_n}\right)\right]^2}{\left[\frac{1}{n}\sum_{i=1}^n \psi_1'\left(\frac{\epsilon_i(\hat{\boldsymbol{\theta}}_{\mathrm{MM}})}{\hat{s}_n}\right)\right]^2}[\boldsymbol{F}^T(\hat{\boldsymbol{\theta}}_{\mathrm{MM}})\boldsymbol{F}(\hat{\boldsymbol{\theta}}_{\mathrm{MM}})]^{-1} \quad \text{com} \quad \boldsymbol{F}(\boldsymbol{\theta}) = \left[\left(\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \theta_j}\right)\right],$$

onde $\boldsymbol{F}$ é a matriz Jacobiana do modelo em ordem aos parâmetros. Contudo, é necessária cautela no uso desta estimativa, uma vez que esta não é robusta (veja-se a este respeito a secção 1.5 na página 8).

2.2.6 Estimadores τ

Uma outra abordagem devida a Yohai e Zamar (1988) que conjuga simultaneamente elevado ponto de rotura e eficiência sob erros Gaussianos superior à dos estimadores LMS e LTS, é a classe de estimadores τ, introduzidos no âmbito da regressão não-linear por Tabatabai e Argyros (1993). As estimativas τ definem-se pela minimização de uma nova estimativa de escala dos resíduos, τ_n, dada por

$$\tau_n^2 = s_n^2 \frac{1}{n}\sum_{i=1}^n \rho_1\left(\frac{\epsilon_i(\boldsymbol{\theta})}{s_n}\right), \tag{2.19}$$

onde s_n é a estimativa M de escala dos resíduos definida por

$$\frac{1}{n}\sum_{i=1}^n \rho_0\left(\frac{\epsilon_i(\boldsymbol{\theta})}{s_n}\right) = b \quad \text{com} \quad b = 0{,}5 \max \rho_0(u). \tag{2.20}$$

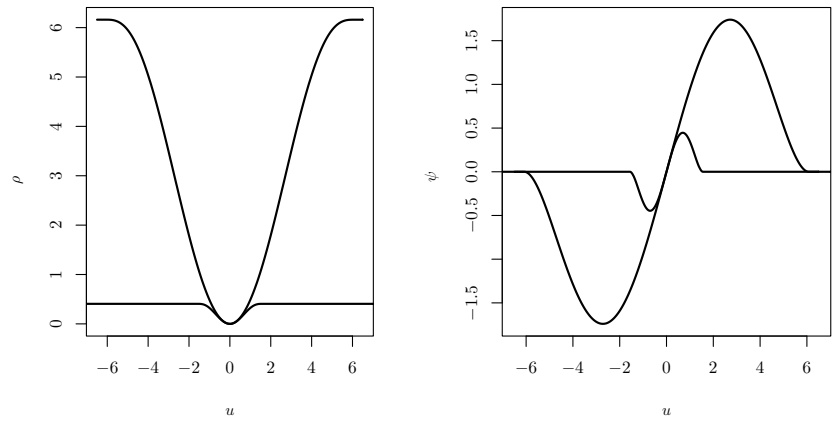


Figura 2.4 Funções ρ usadas na construção do estimador τ e respectivas funções ψ.

A função ρ deve satisfazer propriedades idênticas às referidas para o estimador MM e, adicionalmente, $2\rho_1(u) - \psi_1(u)u \geqslant 0$.

Yohai e Zamar (1988) mostram, para o modelo linear, que o comportamento assimptótico do estimador τ é equivalente ao de um estimador M cuja função ρ seja a média ponderada das duas funções ρ usadas na sua construção. O ponto de rotura depende unicamente de ρ_0, pelo que a escolha de ρ_1 é orientada de forma a obter-se uma estimativa com eficiência elevada quando a distribuição do erro é Gaussiana. Estes autores, bem como Tabatabai e Argyros usam a família de funções dada por

$$\rho_{B,c}(u) = \begin{cases} \frac{u^2}{2}\left(1 - \frac{u^2}{c^2} + \frac{u^4}{3c^4}\right) & \text{se } |u| \leqslant c \\ \frac{c^2}{6} & \text{se } |u| > c. \end{cases}$$

Para $c_0 = 1{,}56$ e $c_1 = 6{,}08$, com $\rho_0 = \rho_{B,c_0}$ e $\rho_1 = \rho_{B,c_1}$, obtêm-se estimativas τ com ponto de rotura 0,5 e eficiência 0,95 sob erros Gaussianos. A figura 2.4 ilustra ρ_0 e ρ_1 bem como as funções de influência respectivas.

2.2.7 Estimador das diferenças aparadas mínimas

Os estimadores MM e τ podem alcançar uma eficiência arbitrária retendo o ponto de rotura de 0,5, mas fazem-lo à custo de um aumento no enviesamento das estimativas (Croux *et al.*, 1994). Uma abordagem alternativa proposta por Stromberg *et al.* (2000) é o estimador das LTD, que por definição minimiza a soma dos menores quadrados das diferenças entre os erros das $\binom{n}{2}$ combinações de pares de observações, isto é,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{k=1}^{\binom{h}{2}} \{(\epsilon_i - \epsilon_j)^2; i < j\}_{k:\binom{n}{2}}, \tag{2.21}$$

onde $\{\,\cdot\,\}_{k:\binom{n}{2}}$ representa a k-ésima estatística de ordem de um conjunto de $\binom{n}{2}$ elementos. A fracção de aparamento, α, é definida de forma idêntica àquela do estimador LTS.

Tabela 2.2 Comparação dos vários tipos de estimadores robustos com resposta univariada

Estimador	Ponto de rotura	Eficiência assimptótica	Postula erro de medição simétrico	Robusto em relação a *outliers* agrupados nas variáveis x
L_p	$1/n$		Sim	Não
M	$1/n$		Sim	Não
LMS	0,5	0%[a]	Sim	Sim
LTS	0,5	7%[a]	Sim	Sim
MM	0,5	80% a 99%[b]	Sim	Sim
τ	0,5	95%[b]	Sim	Sim
LTD	0,5	67%[a]	Não	Sim

[a] No contexto da regressão linear.
[b] Valores para as versões apresentadas neste capítulo.

Convém desde já chamar a atenção para uma particularidade da definição que acabou de introduzir-se. Considere-se o modelo de regressão linear $y = \beta_0 + \beta_1 x$. Neste caso, a função objectivo do estimador LTD não depende do termo independente já que este é cancelado na diferença $\epsilon_i - \epsilon_j$. Consequentemente, β_0 tem de ser estimado separadamente aplicando um estimador de localização a $y - \hat{\beta}_1 x$, em que $\hat{\beta}_1$ designa a estimativa LTD de β_1. Assim, tendo presente a forma do modelo $f(\boldsymbol{x}_i, \boldsymbol{\theta})$, convém não perder de vista a hipótese de ocorrer um fenómeno similar no domínio não-linear. Repare-se que este aspecto constitui uma vantagem se o parâmetro em causa não fizer parte por si só da finalidade da análise. De facto, nesta situação a dimensão do problema de optimização é reduzida relativamente à de outros estimadores em que esse parâmetro é — necessariamente — estimado por "arrasto", dada a sua influência nas estimativas dos restantes.

Uma propriedade deste estimador que decorre do uso das $\binom{n}{2}$ diferenças entre pares de resíduos é acomodar distribuições do ruído de medição assimétricas, uma vez que a distribuição da diferença entre pares $\epsilon_i - \epsilon_j$ é simétrica mesmo quando a distribuição dos ϵ_i o não é (Croux *et al.*, 1994). Com efeito, a função objectivo dos estimadores descritos até aqui atribui peso idêntico indistintamente a erros negativos ou positivos com o mesmo valor absoluto.

Além disso, uma propriedade importante é que as estimativas são resistentes a ruído de medição com componente de erro sistemático. De facto, analogamente à situação do termo independente no modelo de regressão linear, a componente sistemática é cancelada nas diferenças entre pares de erros.

Tal como o estimador LTS, o grau de robustez do estimador LTD aumenta com o aparamento: em regressão linear o valor máximo (0,5) do ponto de rotura é atingido para $h \approx n/2$, para o qual a eficiência assimptótica Gaussiana é 67%.

2.2.8 Sumário das propriedades dos estimadores

A tabela 2.2 apresenta um sumário das principais características dos vários tipos de estimadores robustos acabados de descrever.

2.3 Modelos com resposta multivariada

2.3.1 Estimador de máxima verosimilhança concentrada

Consideremos o modelo de regressão com resposta multivariada que representa n_y respostas (variáveis dependentes) medidas para n conjuntos de valores das variáveis independentes cada uma

$$y_{ij} = f_j(\boldsymbol{x}_i, \boldsymbol{\theta}) + \epsilon_{ij}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, n_y, \tag{2.22}$$

onde y_{ij} designa a j-ésima resposta correspondente à i-ésima observação ou experiência, f_j designa o modelo não-linear para a j-ésima resposta, e ϵ_{ij} designa o erro de medição.

Assumiremos que o vector $\boldsymbol{\epsilon}_i = [\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{in_y}]^T$ tem as seguintes propriedades:

$$E(\boldsymbol{\epsilon}_i) = \mathbf{0} \quad \text{e} \quad E(\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_l^T) = \begin{cases} \boldsymbol{\Sigma} & \text{se } i = l \\ \mathbf{0} & \text{se } i \neq l, \end{cases} \qquad \forall i, l \tag{2.23}$$

onde $\boldsymbol{\Sigma}$ é uma matriz de covariâncias do tipo $n_y \times n_y$ fixa. Por palavras, os erros têm média nula, os erros de experiências diferentes são considerados independentes e assume-se que a matriz de covariâncias é idêntica em cada experiência. Admitindo o pressuposto adicional de que os $\boldsymbol{\epsilon}_i$ seguem a distribuição Gaussiana multivariada, resulta que a função de densidade conjunta das n respostas $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{in_y}]^T$ é dada por

$$(2\pi)^{-nn_y/2} |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{ -\tfrac{1}{2} \sum_{i=1}^{n} [\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\theta})]^T \boldsymbol{\Sigma}^{-1} [\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\theta})] \right\}$$

onde $\boldsymbol{f}_i(\boldsymbol{\theta}) = [f_1(\boldsymbol{x}_i, \boldsymbol{\theta}), f_2(\boldsymbol{x}_i, \boldsymbol{\theta}), \ldots, f_{n_y}(\boldsymbol{x}_i, \boldsymbol{\theta})]^T$.

Deste modo, o logaritmo da função verosimilhança é, a menos duma constante sem importância

$$l = -\frac{n}{2} \ln |\boldsymbol{\Sigma}| - \tfrac{1}{2} \sum_{i=1}^{n} [\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\theta})]^T \boldsymbol{\Sigma}^{-1} [\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\theta})]. \tag{2.24}$$

Logo, a maximização da função log-verosimilhança é equivalente a minimizar (Seber e Wild, 1989, p. 537)

$$\ln |\boldsymbol{\Sigma}| + \operatorname{tr} \{\boldsymbol{\Sigma}^{-1} \boldsymbol{V}(\boldsymbol{\theta})/n\}, \qquad \boldsymbol{V}(\boldsymbol{\theta}) = [\boldsymbol{Y} - \boldsymbol{G}(\boldsymbol{\theta})]^T [\boldsymbol{Y} - \boldsymbol{G}(\boldsymbol{\theta})],$$

onde $\boldsymbol{Y} = [(y_{ij})]$ e $\boldsymbol{G}(\boldsymbol{\theta}) = [(f_j\{\boldsymbol{x}_i, \boldsymbol{\theta}\})]$.

Aplicando a técnica de concentração da verosimilhança, é possível mostrar que $\boldsymbol{\Sigma} = \boldsymbol{V}(\boldsymbol{\theta})/n$ (Bates e Watts, 1988, p. 138), e substituindo na equação anterior, tem-se

$$\ln |\boldsymbol{V}(\boldsymbol{\theta})/n| + n_y. \tag{2.25}$$

Assim, a estimativa de máxima verosimilhança concentrada $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ obtém-se minimizando $|\boldsymbol{V}(\boldsymbol{\theta})|$ e a estimativa ML de $\boldsymbol{\Sigma}$ é dada por

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})/n. \tag{2.26}$$

2.3.2 Estimadores de máxima verosimilhança aparada

O ponto de partida dos estimadores de MTL propostos por Hadi e Luceño (1997) é a substituição da função de verosimilhança do estimador ML por uma versão aparada. Matematicamente

$$\sum_{i=a}^{b} \{\ell(\boldsymbol{\theta})\}_{1:n}, \tag{2.27}$$

onde $\{\ell\}_{1:n} \leqslant \ldots \leqslant \{\ell\}_{n:n}$ representam as contribuições ordenadas de cada observação para a função log-verosimilhança

$$l = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{y}_i) = \sum_{i=1}^{n} \ln g(\boldsymbol{y}_i; \boldsymbol{\theta}),$$

e $a \leqslant b$, $(a, b) \in \{1, 2, \ldots, n\}$, são parâmetros especificados que definem o compromisso entre a robustez e eficiência dos estimadores: valores de a afastados de 1 e de b afastados de n conduzem a estimadores mais robustos mas menos eficientes.

O estimador MTL inclui métodos clássicos e robustos de estimação como casos especiais. Em particular, para modelos com resposta univariada

1. quando $a = 1$ e $b = n$ coincide com o estimador ML;
2. se, adicionalmente, os erros de medição seguirem a distribuição Gaussiana, é o estimador LS;
3. quando $a = 1$, $b < n$, e os erros de medição são Gaussianos, então é o estimador LTS; e
4. quando se maximiza $\operatorname{med}_i \ell(\boldsymbol{\theta}; \boldsymbol{y}_i)$, o estimador resultante denomina-se MML, e corresponde ao estimador LMS se os erros de medição forem Gaussianos.

Em face do que acaba de expor-se, torna-se natural a extensão ao caso de resposta multivariada dos estimadores LMS e LTS combinando os estimadores MTL com a distribuição Gaussiana multivariada. Assim, atendendo à expressão (2.24), tem-se

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}_i) = -\tfrac{1}{2} \ln |\boldsymbol{\Sigma}| - \tfrac{1}{2} [\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\theta})]^T \boldsymbol{\Sigma}^{-1} [\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\theta})].$$

Segue-se que o estimador MML é dado por

$$\begin{aligned} \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}, \boldsymbol{\Sigma}} \quad & \operatorname*{med}_i \{\ln |\boldsymbol{\Sigma}| + [\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\theta})]^T \boldsymbol{\Sigma}^{-1} [\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\theta})]\} \\ \text{sujeito a} \quad & \boldsymbol{\Sigma} \text{ é definida positiva,} \end{aligned} \tag{2.28}$$

e o estimador MTL com $a = 1$ e $b = h$

$$\begin{aligned} \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}, \boldsymbol{\Sigma}} \quad & h \ln |\boldsymbol{\Sigma}| + \sum_{j=1}^{h} \{[\boldsymbol{y}_{i_j} - \boldsymbol{f}_{i_j}(\boldsymbol{\theta})]^T \boldsymbol{\Sigma}^{-1} [\boldsymbol{y}_{i_j} - \boldsymbol{f}_{i_j}(\boldsymbol{\theta})]\}_{j:n} \\ \text{sujeito a} \quad & \boldsymbol{\Sigma} \text{ é definida positiva,} \end{aligned} \tag{2.29}$$

onde $i_j \in J$ designa o índice das h observações com menor valor da distância de Mahalanobis. A dimensão do problema de optimização correspondente é $n_p + n_y(n_y+1)/2$.

Se em cada experiência considerarmos que os erros das várias respostas são independentes, isto é, a matriz de covariâncias $\boldsymbol{\Sigma}$ é diagonal, então, a partir das equações (2.28) e (2.29) podemos escrever

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\text{MML}} = \arg\min_{\boldsymbol{\theta},\boldsymbol{\sigma}} \quad & 2\ln\prod_{j=1}^{n_y}\sigma_j + \operatorname*{med}_i\left\{\sum_{j=1}^{n_y}\left(\frac{y_{ij}-f_{ij}(\boldsymbol{\theta})}{\sigma_j}\right)^2\right\} \\ \text{sujeito a} \quad & \sigma_j > 0, \end{aligned} \tag{2.30}$$

e

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\text{MTL}} = \arg\min_{\boldsymbol{\theta},\boldsymbol{\sigma}} \quad & 2h\ln\prod_{j=1}^{n_y}\sigma_j + \sum_{i=1}^{h}\left\{\sum_{j=1}^{n_y}\left(\frac{y_{ij}-f_{ij}(\boldsymbol{\theta})}{\sigma_j}\right)^2\right\}_{i:n} \\ \text{sujeito a} \quad & \sigma_j > 0. \end{aligned} \tag{2.31}$$

Por uma questão de simplicidade computacional o presente estudo limita-se a este caso particular (veja-se a secção 2.4.3 na página 37).

Um aspecto interessante a realçar é que o estimador MTL é essencialmente equivalente à robustificação do estimador de máxima verosimilhança concentrada se se recorrer ao estimador determinante da covariância mínimo (MCD). Para se chegar a este resultado, passamos a descrever a ideia geral subjacente a esta técnica e referem-se sucintamente os aspectos básicos do estimador MCD — um estimador de localização e dispersão multivariado com elevado ponto de rotura.

Como se viu, a determinação da matriz de covariâncias clássica dos resíduos, $\boldsymbol{V}/n$, é pedra angular do estimador de máxima verosimilhança concentrada multivariado. Seguindo Rousseeuw e Leroy (1987, p. 269 e 270), uma forma de o robustificar consiste em substituir $\boldsymbol{V}$ por uma estimativa robusta de dispersão multivariada.

O estimador MCD identifica inicialmente a subamostra com h observações diferentes indexada por $J = \{i_1, i_2, \ldots, i_h\}$ para a qual o determinante da matriz de covariâncias convencional dos erros

$$\boldsymbol{C} = \frac{1}{h-1}\sum_{i\in J}(\boldsymbol{\epsilon}_i - \bar{\boldsymbol{\epsilon}}_J)(\boldsymbol{\epsilon}_i - \bar{\boldsymbol{\epsilon}}_J)^T, \quad \text{onde} \quad \bar{\boldsymbol{\epsilon}}_J = \frac{1}{h}\sum_{i\in J}\boldsymbol{\epsilon}_i, \tag{2.32}$$

é mínimo. O resultado anterior é refinado seleccionando os pontos cuja distância de Mahalanobis não é demasiado grande (Venables e Ripley, 1999, p. 348), mais precisamente,

$$J' = \{\, i : n(\boldsymbol{\epsilon}_i - \bar{\boldsymbol{\epsilon}}_J)^T\boldsymbol{C}^{-1}(\boldsymbol{\epsilon}_i - \bar{\boldsymbol{\epsilon}}_J) \leqslant \chi^2_{n_y}(0{,}025),\ i \in J\,\}.$$

A estimativa da matriz de covariâncias do estimador MCD é então a matriz $\boldsymbol{C}$ com base em J' multiplicada por uma constante de modo a garantir consistência sob a distribuição Gaussiana. Usando $h = \lfloor (n + n_y + 1)/2 \rfloor$ o estimador MCD salvaguarda até 50% de *outliers*. Contudo, se existir evidência que a fracção de *outliers* na amostra é inferior a $\alpha \leqslant 0{,}5$, Rousseeuw e Leroy (1987, p. 263) sugerem o uso de $h = \lfloor n(1-\alpha) \rfloor + 1$, para o qual o ponto de rotura assimptótico do estimador é α.

Em conclusão, a estimativa de máxima verosimilhança concentrada robustificada é dada por

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} |\boldsymbol{C}(\boldsymbol{\theta})|. \tag{2.33}$$

Retornando agora à expressão (2.29), observe-se que esta pode ser posta na forma

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}, \boldsymbol{\Sigma}, J} \left\{ h \ln |\boldsymbol{\Sigma}| + \sum_{i \in J} [\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\theta})]^T \boldsymbol{\Sigma}^{-1} [\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\theta})] \right\}.$$

Para J fixo, esta expressão é análoga à maximização de (2.24), pelo que, de modo semelhante, obtém-se

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ \min_{J} \det \left(\frac{1}{h} \sum_{i \in J} \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T \right) \right\} = \arg\min_{\boldsymbol{\theta}} |\boldsymbol{C'}(\boldsymbol{\theta})|.$$

Comparando as matrizes $\boldsymbol{C'}$ e $\boldsymbol{C}$, repare-se que as diferenças consistem no facto de em $\boldsymbol{C}$ os $\boldsymbol{\epsilon}_i$ serem centrados pela sua média e de excluir-se dos h pontos iniciais aqueles com distâncias de Mahalanobis grandes. Por isso, os estimadores são apenas aproximadamente equivalentes. Contudo, é razoável supor situações em que estes apresentem um comportamento basicamente idêntico, caso em que podem ser encarados como algoritmos diferentes de computação do mesmo estimador.

2.3.3 Estimadores M multivariados

A extensão directa dos estimadores M ao caso multivariado pode fazer-se recorrendo à expressão (2.12) na seguinte forma

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{j=1}^{n_y} \rho \left(\frac{\epsilon_{ij}(\boldsymbol{\theta})}{\sigma_j} \right), \tag{2.34}$$

onde σ_j é um factor de escala correspondente à j-ésima resposta. Note-se que esta formulação ignora a possível correlação das variáveis dependentes.

Koenker e Portnoy (1990) designam os estimadores definidos por (2.34) estimadores M ordinários. Nesse artigo, no âmbito da regressão linear, é proposto um estimador assimptoticamente mais eficiente. Este, é obtido pela modificação das condições de estacionaridade de (2.34), que resultam num sistema de equações que não corresponde a um problema de optimização. Deste modo, o estimador é definido pela solução de um sistema de equações algébricas, e, tal como os estimadores M ordinários não considera a correlação das respostas.

Em contraponto aos estimadores que acabámos de referir, Krishnakumar e Ronchetti (1997) descrevem um estimador que considera a natureza multivariada do problema de forma completa.

Neste estudo simulam-se os estimadores M ordinários. Para calcular os σ_j usa-se a proposta 2 de Huber, e toma-se para ρ a função de Huber com $c = 1{,}5$. Deste modo, a partir da equação (2.34) podemos escrever

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^{n_y} \sigma_j \left[\sum_{i=1}^{n} \rho_H \left(\frac{\epsilon_{ij}(\boldsymbol{\theta})}{\sigma_j} \right) + a \right], \tag{2.35}$$

onde ρ_H representa a função de Huber e $a = (n - n_p) E_\Phi(\psi_H^2/2)$.

2.3.4 Estimador LAD multivariado (norma ℓ_1 de ℓ_2)

Considere-se a norma ℓ_v de ℓ_w de um vector $\boldsymbol{u}$, definida por

$$\|u\| = \left(\sum_i |u_i|^w\right)^{v/w}.$$

No contexto do modelo de regressão linear multivariada, Kaufman *et al.* (2002) descrevem o estimador LAD com referência à norma ℓ_1 de ℓ_2. Como o nome indica, este método minimiza a soma da norma dos erros $\boldsymbol{\epsilon}_i$ com $v = 1$ e $w = 2$. Tem-se, assim,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(\sum_{j=1}^{n_y} |\epsilon_{ij}|^2\right)^{1/2}. \tag{2.36}$$

Parece razoável supor que o comportamento deste estimador seja intermédio entre o do método dos mínimos quadrados (regressão ℓ_2 de ℓ_2) e o do estimador obtido para $v = 1$ e $w = 1$, que minimiza

$$\sum_{i=1}^{n}\sum_{j=1}^{n_y} |\epsilon_{ij}|.$$

De facto, o estimador LAD incorpora elementos de ambos os métodos. Note-se que quando $n_y = 1$, a regressão ℓ_1 de ℓ_2 é idêntica à regressão ℓ_1 de ℓ_1 (neste caso é o estimador LAD univariado), e se $n = 1$ é idêntica ao estimador LS.

2.4 Aspectos computacionais

A discussão que se segue trata de questões dos procedimentos de cálculo dos estimadores descritos nas secções precedentes. Começamos com uma nota sobre os algoritmos numéricos de regressão robusta disponíveis na literatura para o caso não-linear.

2.4.1 Algoritmos para regressão não-linear robusta existentes

Estimadores LMS e LTS

As respectivas funções objectivo são contínuas, não diferenciáveis, e não convexas, apresentando múltiplos mínimos locais. Consideremos, sem perda de generalidade, o estimador LTS. Uma abordagem exacta para o seu cálculo, consiste na enumeração exaustiva da totalidade dos $\binom{n}{h}$ subconjuntos com h elementos da amostra original, e para cada subamostra resolver um problema de mínimos quadrados. A estimativa LTS corresponde à estimativa LS com o menor valor da soma do quadrado dos resíduos. No entanto, esta abordagem apenas é praticável em amostras de pequena dimensão, caso em que o número de subconjuntos a considerar não conduz a uma carga computacional proibitiva.

No âmbito da regressão linear, a abordagem mais comum para contornar esta limitação é a técnica de reamostragem de subamostras de n_p elementos de Rousseeuw e Leroy (1987), implementada no algoritmo PROGRESS. A ideia por detrás é examinar um número de subconjuntos de cardinal idêntico ao número de parâmetros do modelo (subconjuntos-n_p), escolhido de forma a assegurar, com elevada probabilidade (digamos

0,999), a presença de pelo menos uma subamostra sem *outliers*. Para cada subamostra é calculado o ajuste exacto — que consiste, neste caso, na solução de um sistema linear de n_p equações a n_p incógnitas —, e a estimativa assim obtida é usada no cálculo do respectivo valor da função objectivo na amostra original. A estimativa LTS corresponde aquela com o menor valor da função objectivo.

Uma outra abordagem consiste no algoritmo de permuta de Hawkins (1994). Neste método, parte-se de um conjunto arbitrário de h elementos (subconjuntos-h) e resolve-se o correspondente problema de mínimos quadrados. Realizando sucessivas permutas entre pares de observações — uma do subconjunto corrente e outra das $n - h$ observações não seleccionadas —, que resultem na diminuição do valor da função objectivo, refina-se a estimativa, a qual converge para um extremo local. Repetindo o procedimento para outros subconjuntos iniciais gerados aleatoriamente, o método conduz, com probabilidade arbitrariamente elevada, ao extremo global à medida que se aumenta o número de conjuntos iniciais.

Tendo presente que no caso não-linear o esforço computacional de um problema de mínimos quadrados ou de solução de um sistema de equações algébricas (não-lineares) é significativamente maior que no caso linear, por um lado, e que o número de subconjuntos examinado na prática, no caso linear, é elevado em ambas as técnicas, por outro, concluise que a aplicação directa destas técnicas ao caso não-linear conduziria a um tempo de cálculo excessivo.

Stromberg (1993) apresenta um algoritmo — designado por MSA (*multistage algorithm*) em Chen *et al.* (1997) — para o estimador LMS no caso não-linear,[6] que inclui a técnica de reamostragem de subconjuntos-n_p.

Mais recentemente, Chen *et al.* (1997) desenvolveram o algoritmo SHA (*six half algorithm*) que se inspira na técnica de permuta, e mostram que o seu desempenho numérico é superior ao do algoritmo anterior. Tanto num caso como noutro reduziu-se o número de subconjuntos a examinar no caso não-linear a um valor razoável. O problema de mínimos quadrados é resolvido por Stromberg e Chen *et al.* pelo método de Newton-Raphson e Gauss-Newton, respectivamente.

Estimadores M e τ

Da literatura sobre estimadores M destacamos os algoritmos apresentados por Dutter e Huber (1981) e Edlund *et al.* (1997). Tabatabai e Argyros (1993) desenvolveram um algoritmo para os estimadores τ. Os métodos referidos são algoritmos de optimização determinísticos (gradiente) de procura local, isto é, apenas garantem a convergência para uma solução óptima local.

2.4.2 Escolha do algoritmo de optimização

Em nosso entender, as características especiais dos problemas de estimação não-linear de parâmetros e do método de Monte Carlo tornam particularmente inadequada a utilização dos métodos de optimização atrás referidos. Vejamos porquê:

[6] O algoritmo é igualmente aplicável ao estimador LTS.

- Frequentemente, não se dispõe duma ideia da grandeza dos valores dos parâmetros do modelo matemático. É bem conhecido o melindre e dificuldade dum procedimento de cálculo de estimativas iniciais, cuja natureza é específica do problema em estudo. Embora existam na literatura linhas de orientação gerais e recomendações de procedimentos para alguns casos particulares (veja-se, por exemplo, Bates e Watts, 1988, secção 3.3), não parece ser possível o desenvolvimento dum procedimento genérico para atacar este problema.

- A natureza não-linear da generalidade dos sistemas em engenharia química traduz-se frequentemente, mesmo quando o critério de regressão associado é convexo, num problema de optimização não-convexo, que pode apresentar vários óptimos locais (Esposito e Floudas, 1998). É claro que as diferenças entre as soluções globais e locais podem ser extremamente significativas.

Fica assim patente que um algoritmo clássico baseado em informação das derivadas da função objectivo não é o mais adequado no âmbito de uma simulação de Monte Carlo. Com efeito, estes algoritmos convergem para soluções locais e o desenrolar do processo iterativo é extremamente sensível à qualidade das soluções iniciais. O procedimento vulgarmente seguido para contornar este problema consiste no uso de vários pontos de partida diferentes. Ora, se recordarmos que uma experiência de Monte Carlo exige tipicamente um número elevado de réplicas, reconhece-se facilmente que o esforço computacional associado a esta estratégia é incomportável. Note-se que o facto de dispormos do "verdadeiro" valor dos parâmetros não significa, necessariamente, dispormos de um ponto de partida razoavelmente próximo da solução, porque a estimativa associada com uma amostra particular pode estar muito afastada do verdadeiro valor do parâmetro. Por outro lado, se o modelo matemático não tem forma explícita, torna-se necessário um procedimento computacional de simulação. Eventualmente, características particulares do procedimento computacional (*e.g.*, inclusão de algoritmos adaptativos, condições de paragem, lógica se-então-senão) acoplam ruído numérico aos valores calculados. Isto pode dar origem a erros graves no cálculo das derivadas da função objectivo (Kolda *et al.*, 2003, pp. 391 e 392) usando diferenças finitas ou diferenciação automática. Nesta situação, estes autores recomendam a utilização de algoritmos directos[7] (veja-se também, Kelley, 1999, p. 111).

- A questão central que se levanta num quadro de comparação empírica entre estimadores consiste em perceber a importância do efeito do algoritmo de optimização a utilizar num dado estimador sobre o desempenho desse estimador. Uma solução simples que evita este problema é usar o mesmo algoritmo para os diversos estimadores. Contudo, isto requer um método de optimização bastante versátil para acomodar o maior número possível de classes de estimadores.

Sendo assim, é fácil de ver que os métodos gradiente têm uma gama de aplicabilidade limitada na medida em que se circunscrevem apenas a estimadores baseados em funções objectivo diferenciáveis.

[7] Entendido num sentido geral como aqueles que usam apenas a informação dos valores da função objectivo.

Em conclusão, os métodos apresentados na secção anterior não permitiriam ter em conta as questões acabadas de apontar.

Por conseguinte, advoga-se o uso de algoritmos directos de carácter global. De acordo com Pínter (2002, pp. 518 e 519) entre as inúmeras classes da taxonomia dos métodos de optimização global, a mais eficiente à luz do tempo de cálculo é possivelmente a das heurísticas estocásticas. Embora não garantam a localização do óptimo global ao contrário dos procedimentos determinísticos, em contrapartida a sua maior eficiência computacional é absolutamente essencial num contexto de experimentação de Monte Carlo.

Uma vantagem adicional das técnicas heurísticas é a sua generalidade, isto é, a insensibilidade aos detalhes da estrutura do problema, o que possibilita o uso de um único algoritmo para o cálculo dos diferentes estimadores apresentados nas secções precedentes. Deste modo, a análise dos resultados de Monte Carlo não será obscurecida pelo comportamento numérico de métodos diferentes.

Para este trabalho consideraram-se os seguintes métodos: pesquisa aleatória controlada modificada (MCRS) e um algoritmo de pesquisa evolutiva (ES2), ambos propostos por Křivý *et al.* (2000), evolução diferencial (DE), introduzido por Storn e Price (1997); Price (1999), e uma variante deste último, designada MDE, desenvolvida por Lee *et al.* (1999) com o objectivo de acelerar a velocidade de convergência. As técnicas referidas enquadram-se na vasta área da evolução computacional, que se inspira num paradigma biológico. Em termos muito gerais, estas técnicas baseiam-se num procedimento iterativo que faz evoluir um conjunto ou *população* de soluções candidatas por intermédio da aplicação de operadores que emulam os processos de evolução biológicos. Experiências preliminares indicaram que o algoritmo MDE apresenta a melhor combinação de fiabilidade e eficiência, tendo este sido usado na totalidade das simulações deste estudo. Na figura 2.5 na página seguinte apresenta-se o algoritmo MDE.

2.4.3 Implementação

Os estimadores (e as experiências de Monte Carlo) foram programados na linguagem S (Becker *et al.*, 1988; Chambers e Hastie, 1992; Chambers, 1998) sendo usada a implementação do sistema R (Ihaka e Gentleman, 1996). O método de geração de números aleatórios usado é o método Marsaglia-Multicarry, sendo o controlo das sementes estabelecido pela função `set.seed`

População inicial

A população inicial é gerada aleatoriamente com uma distribuição uniforme numa região hiperrectangular, sendo possível incluir indivíduos (pontos do espaço dos parâmetros) que à partida se julga constituírem estimativas razoáveis da solução, com o objectivo de acelerar a localização do óptimo global (Cela *et al.*, 2001).

Restrições sobre os parâmetros

Um método possível de optimizar uma função cuja computação envolve a resolução dum modelo matemático consiste em obter uma solução numérica desse modelo sempre que o algoritmo de optimização necessite de avaliar o valor da função objectivo. Numa outra

Inicialização: Escolher o número máximo de gerações, o tamanho da população, $N_{\mathrm{P}} \geqslant 3$ ($10d$, d denota o número de variáveis de decisão), o factor de ponderação base, F_{b} (0,5), o factor de *crossover* $C_{\mathrm{R}} \in [0,1]$ (0,8), e o passo $\xi \in [1{,}3, 1{,}7]$ (1,5). (Entre parêntesis indicam-se os valores por omissão usados.)
Estipular a população inicial $\boldsymbol{z}_i = [z_{1i}, z_{2i}, \ldots, z_{di}]^T, \quad i = 1, \ldots, N_{\mathrm{P}}$, onde cada indivíduo $\boldsymbol{z}_i$ representa um ponto do espaço das variáveis de decisão.
repetir
  $F = F_{\mathrm{b}}$
  incrementar o contador de gerações
  para $i = 1, \ldots, N_{\mathrm{P}}$ **fazer**
    Na G-ésima geração escolhe-se o melhor indivíduo, $\boldsymbol{z}_{\mathrm{b}}$, e seleccionam-se aleatoriamente os índices r_1 e r_2 do conjunto $\{1, 2, \ldots, N_{\mathrm{P}}\}$ tais que $i \neq r_1 \neq r_2$. Tem-se, então,

$$\hat{\boldsymbol{z}}_i^{G+1} = \boldsymbol{z}_i^G + F(\boldsymbol{z}_{\mathrm{b}}^G - \boldsymbol{z}_i^G) + F(\boldsymbol{z}_{r_1}^G - \boldsymbol{z}_{r_2}^G).$$

    (Operação de mutação)
    para $j = 1, \ldots, d$ **fazer**
      A descendência $\boldsymbol{z}_i^{G+1} = [z_{1i}^{G+1}, z_{2i}^{G+1}, \ldots, z_{di}^{G+1}]^T$ obtém-se a partir de
      se $\delta_{ji} \geqslant C_{\mathrm{R}}$ **então** (Operação de *crossover*)
        $z_{ji}^{G+1} = z_{ji}^G$
      se não
        $z_{ji}^{G+1} = \hat{z}_{ji}^{G+1}$
      fim (se)
      onde $\delta_{ji} \in [0,1]$ é um número aleatório que segue a distribuição uniforme. Note-se que o valor de C_{R} determina a diversidade da descendência gerada, uma vez que quanto mais elevado for maior é a probabilidade do elemento $\hat{z}_{ji}^{G+1}$ do novo indivíduo incorporar a geração seguinte.
    fim (para)
    Verificar se os indivíduos da nova população se encontram dentro dos limites dos parâmetros. No caso de um elemento de um indivíduo violar as fronteiras, tomar uma medida correctiva.
    repetir (Pesquisa local)
      $F = \xi F$
      Gerar um indivíduo $\boldsymbol{z}_i^{G+1}$ novo, com o novo F e as operações de mutação e *crossover* apresentadas acima. Os valores de $\boldsymbol{z}_i^G$, $\boldsymbol{z}_{\mathrm{b}}^G$, $\boldsymbol{z}_{r_1}^G$, e $\boldsymbol{z}_{r_2}^G$ mantêm-se.
      Terminar quando (suponhamos o problema de minimização) $\boldsymbol{z}_i^{G+1}$ não diminuir o valor da função objectivo J.
    fim (repetir)
    se $J(\boldsymbol{z}_i^{G+1}) < J(\boldsymbol{z}_i^G)$ **então** (Selecção dos indivíduos da nova geração)
      $\boldsymbol{z}_i^{G+1} = \boldsymbol{z}_i^{G+1}$
    se não
      $\boldsymbol{z}_i^{G+1} = \boldsymbol{z}_i^G$
    fim (se)
  fim (para)
  Terminar quando o critério de paragem for satisfeito ou se atingir o número máximo de gerações.
fim (repetir)

Figura 2.5 Algoritmo MDE (adaptado de Lee *et al.* 1999).

abordagem consiste os processos de optimização e de resolução do modelo matemático decorrem simultaneamente. O primeiro método apresenta a vantagem relativamente ao segundo de que a sua implementação computacional é directa e, por isso, é o utilizado neste trabalho. Uma fraqueza associada a esta abordagem é que o algoritmo de optimização pode gerar valores para os quais a resolução numérica do modelo matemático falhe, o que geralmente interrompe o processo de cálculo.

Por conseguinte, é necessário que o algoritmo de optimização incorpore restrições aos valores dos parâmetros, de modo que a pesquisa se mantenha na região admissível "de facto". Uma segunda razão decorre do significado físico dos parâmetros, que, normalmente, implica uma gama limitada de valores possíveis. Tem-se assim,

$$\begin{bmatrix} \boldsymbol{\theta}^L \\ \boldsymbol{\gamma}^L \end{bmatrix} \leqslant \begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{\gamma} \end{bmatrix} \leqslant \begin{bmatrix} \boldsymbol{\theta}^U \\ \boldsymbol{\gamma}^U \end{bmatrix}.$$

onde o vector $\boldsymbol{\gamma}$ (se existir) serve de veículo a parâmetros incómodos (frequentemente designados por parâmetros perturbadores) que devem ser estimados juntamente com os parâmetros de interesse do modelo, e os expoentes L e U designam o limite inferior e superior, respectivamente.

Sem dúvida que a técnica que mais vulgarmente aparece na literatura das heurísticas evolutivas é a reposição na fronteira do domínio admissível das componentes de soluções candidatas que ultrapassem os limites. Precisando para a componente (parâmetro) j do indivíduo i da população de soluções:

$$z_{ji}^{G+1} = \begin{cases} \theta_j^L & \text{se } z_{ji}^{G+1} < \theta_j^L \\ \theta_j^U & \text{se } z_{ji}^{G+1} > \theta_j^U. \end{cases}$$

Muito embora esta técnica seja apropriada na prática, Price (1999, p. 86) repara que pode diminuir a diversidade da população, e propõe em alternativa situar a componente transgressora num ponto intermédio entre o limite violado e valor da geração anterior, isto é,

$$z_{ji}^{G+1} = \begin{cases} (\theta_j^L + z_{ji}^G)/2 & \text{se } z_{ji}^{G+1} < \theta_j^L \\ (\theta_j^U + z_{ji}^G)/2 & \text{se } z_{ji}^{G+1} > \theta_j^U, \end{cases} \tag{2.37}$$

que é o método usado neste trabalho, apesar de os resultados obtidos em algumas experiências informais indiciarem que, do ponto de vista prático, os dois procedimentos não se revelam significativamente diferentes um do outro.

Critérios de paragem

Cada iteração de um método evolutivo gera um conjunto de indivíduos (pontos no espaço dos parâmetros), ao contrário de um algoritmo determinístico clássico que gera apenas um ponto. Logo, a natureza dos critérios de convergência dos primeiros é inteiramente distinta dos usualmente empregues nestes últimos.

Tendo presente que num método evolutivo a variabilidade das sucessivas gerações diminui — embora num padrão não necessariamente monótono —, é natural utilizar uma medida da variabilidade de uma geração como critério de convergência. Křivý e

Tvrdík (1995, 1999); Křivý *et al.* (2000) apresentam um critério para regressão com modelos de resposta univariada, definido como,

$$\frac{f(\boldsymbol{z}_k^G) - f(\boldsymbol{z}_{\min}^G)}{f_0} \leqslant \epsilon_0, \tag{2.38}$$

em que ε_0 é o valor da tolerância, o índice min designa o vector de parâmetros da população com o menor valor da função objectivo, $\boldsymbol{z}_k^G$ representa outro ponto da população — por, exemplo, aquele com o valor máximo da função objectivo —, e f_0 é uma constante determinada pela variabilidade da variável dependente.

Estes autores propõem para valor de f_0 a soma do quadrados dos desvios entre a variável dependente e a respectiva média, ou seja,

$$f_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Em experiências numéricas observámos que as populações geradas pelo algoritmos DE e MDE apresentam a partir de determinada altura uma pequena fracção de indivíduos que não evoluem, isto é, existem z_{ji}^G que se mantêm constantes.

Este comportamento, exclui o uso de $\boldsymbol{z}_{\max}^G$ para $\boldsymbol{z}_k^G$ na expressão 2.38, visto que este irá coincidir necessariamente com um dos indivíduos que não evolui, o que impede o decréscimo do valor do critério e, consequentemente, a verificação da condição de paragem. Assim, propomos a utilização da mediana da população, valor que segundo a nossa experiência conduz às soluções apropriadas.

Para estender o critério de convergência apresentado ao caso com modelos de resposta multivariada apenas se torna necessária redefinir f_0. Propomos a utilização das seguintes expressões para f_0:

Estimador ML concentrado e ML robustificado:	$\det(\boldsymbol{Z}^T\boldsymbol{Z})$,
Estimadores MML e MTL:	$\sum_i \sum_j (y_{ij} - \bar{y}_j)^2$,
Estimador M (proposta 2 de Huber):	$\sum_j \hat{\sigma}_j \left[\sum_i \rho\left(\frac{y_{ij} - \bar{y}_j}{\hat{\sigma}_j} \right) + a \right]$,
Estimador LAD (norma ℓ_1 de ℓ_2):	$\sum_i \left[\sum_j (y_{ij} - \bar{y}_j)^2 \right]^{1/2}$,

onde $\boldsymbol{Z} = [(y_{ij} - \bar{y}_j)]$, $\bar{y}_j = \sum_i y_{ij}/n$, e $\hat{\sigma}_j = 1{,}4826\,\mathrm{med}_i|y_{ij} - \bar{y}_j|$.

Ordenação

Da definição dos estimadores aparados apresentados neste capítulo decorre a ordenação de diferentes quantidades: dos quadrados dos resíduos nos estimadores LMS e LTS, dos quadrados das diferenças entre pares de resíduos no estimador LTD, ou das log-verosimilhanças ℓ_i nos estimadores MTL e MML.

Note-se que é suficiente definir a lista de elementos com valor inferior ou igual ao valor do $(1-\alpha)$-ésimo quantil, não sendo necessário que os elementos deste conjunto se encontrem ordenados.

Assim, deve-se efectuar apenas a ordenação *parcial* do conjunto, dado o esforço de cálculo adicional desnecessário que decorre duma ordenação completa. Este aspecto é crucial para a eficiência computacional da implementação destes estimadores.

Estimador MM

Para o cálculo da estimativa de escala s_n, usámos a função `uniroot` — que utiliza o algoritmo descrito em Brent (1973, capítulo 4) — na resolução da equação algébrica não-linear (2.17). O intervalo inicial de pesquisa é $[0,\ \mathrm{MAD}(\hat{\epsilon}_i)]$, em que $\hat{\epsilon}_i$ designa os resíduos do estimador de elevado ponto de rotura inicial. O limite superior foi determinado empiricamente.

Como já referimos, na minimização de (2.18) — a qual determina a estimativa final dos parâmetros —, é suficiente garantir a convergência para um óptimo local. Por outro lado, vimos anteriormente que é vantajoso utilizar um algoritmo que permita restringir o domínio de procura. Assim, optámos por usar o algoritmo de procura local L-BFGS-B (Byrd *et al.*, 1995).

Estimadores τ

O cálculo dos estimadores τ traduz-se num problema de optimização com restrições, em que a equação (2.20) que define a estimativa M de escala dos resíduos, s_n, é uma restrição de igualdade. Infelizmente, as técnicas de incorporação de restrições nos algoritmos evolutivos apresentam inúmeras dificuldades práticas (Coello, 2002):

- Um problema importante é a larga variedade de abordagens existentes, cada uma com os seus prós e contras; Coello apresenta algumas recomendações genéricas, mas a escolha é basicamente um processo de tentativa e erro.
- Outro é que o desempenho de muitos algoritmos é altamente dependente dos valores dos respectivos parâmetros de controlo; portanto, é crucial a escolha cuidadosa destes valores, o que é geralmente um procedimento difícil e moroso.
- Uma terceira desvantagem é o custo computacional severo de algumas abordagens.

Por outro lado, o problema de resolução da equação algébrica não-linear (2.20) é bem colocado, dispondo-se de uma estimativa inicial de boa qualidade dada por $s_n^0 = 1{,}4826\,\mathrm{med}_i|\epsilon_i|$ (Rousseeuw e Leroy, 1987, p. 174).

Deste modo, optamos por formular o procedimento de estimação como um problema sem restrições. O preço a pagar é o esforço computacional de solução de s_n associado à avaliação da função objectivo. O cálculo de s_n é igualmente formulado como um problema de optimização, minimizando-se o quadrado da diferença dos membros da equação (2.20) por um método de tipo Newton apresentado em Schnabel *et al.* (1985) usado pela função `nlm`.

Estimadores MTL e MML

Como já foi dito, nestes estimadores optámos por restringir a estimação dos elementos da matriz de covariâncias $\mathbf{\Sigma}$ às variâncias. Uma razão prende-se com as questões referidas

na secção anterior relativas à resolução de problemas com restrições por algoritmos evolutivos. Ora, com esta limitação é suficiente impor condições de não negatividade às variâncias, o que, como se viu na secção 2.4.3 na página 33, é de fácil implementação no âmbito dos algoritmos evolutivos.

Uma segunda razão é que $n_y(n_y + 1)/2$ parâmetros adicionais a estimar implicam a utilização de tamanhos de população que incorrem num custo computacional incompatível com a realização das simulações num espaço de tempo razoável, dado o poder de cálculo disponível.

Para terminar, refira-se que o espaço de procura por omissão para as variâncias é

$$1 \times 10^{-10} \leqslant \sigma_j^2 \leqslant 1{,}4826 \operatorname*{med}_i |y_{ij} - \operatorname*{med}_i y_{ij}|, \qquad j = 1, 2, \ldots, n_y.$$

Capítulo 3

Uma visão geral das experiências de Monte Carlo

3.1 Estudos de simulação existentes

Na literatura estatística, Stromberg (1993) apresenta um estudo de simulação em que compara os estimadores LMS e MM para três modelos diferentes: o modelo exponencial $y = \exp(\theta_0 + \theta_1 x)$, o modelo de Michaelis-Menten com a parametrização $y = \theta_0 x/(\exp\theta_1 + x)$, e o modelo de isomerização descrito por Carr (1960). Neste último, ao contrário deste trabalho, não é utilizado o planeamento de experiências original, sendo os pontos no espaço das variáveis independentes distribuídos uniformemente na região definida por Carr. O ruído de medição é gerado pela distribuição Gaussiana, com 40% e sem *outliers* na variável dependente. No caso do modelo exponencial é investigado o efeito da presença de pontos de repercussão. O estimador MM apresenta um desempenho bom, mas o comportamento do estimador LMS é pobre.

Mais recentemente, Midi (1999) apresenta um estudo de simulação artificial do estimador M de Huber com estimador preliminar para estimativa do factor de escala (veja-se a discussão na página 19). A ideia principal deste artigo é o estudo da influência do estimador preliminar. Para tal considera o estimador LAD obtido por três algoritmos numéricos diferentes e o método dos mínimos quadrados no papel de estimador preliminar. São efectuadas simulações com dados sem *outliers*, dados com um ponto de repercussão, e dados com um *outlier* na resposta, para os modelos de Gompertz, logístico, e de Ricker, definidos por,

$$y = 6\exp[-\exp(0{,}7 - 0{,}4x)], \quad y = 20[1 + \exp(3 - 0{,}5x)]^{-1}, \quad \text{e} \quad y = 2x_i \exp(-0{,}04x).$$

Os resultados, bem como vantagens do ponto de vista computacional destacam o estimador LAD calculado pelo método de Koenker e Park (1992).

No campo da quimiometria, Tan *et al.* (1999) usando dados simulados e reais, investigam o comportamento dos estimadores LS e LTS com 50% de aparamento, concluindo pela superioridade deste último. Nas simulações o erro de medição é gerado pela distribuição Gaussiana e Gaussiana contaminada (são considerados casos com distribuição contaminante de média não nula), com e sem *outliers* na variável dependente.

Por último, Wang *et al.* (2000) descrevem o estimador M robusto adaptativo baseado em ôndulas (WARME), que é essencialmente um estimador de máxima verosimilhança onde em alternativa à pressuposição de uma distribuição para o erro de medição, se estima adaptativamente a respectiva função densidade de probabilidade a partir dos resíduos por um procedimento não-paramétrico baseado em ôndulas. O comportamento

deste estimador é estudado na simulação de um caso artificial que considera uma reacção isotérmica reversível do tipo A $\rightleftarrows$ B a decorrer num reactor tanque contínuo com agitação (CSTR), e é comparado com os estimadores LS, M de Huber, e estimador M robusto adaptativo para identificação não-paramétrica de sistemas (ARMENSI) de Wu e Çinar (1996) — este último pode ser interpretado como a solução do procedimento apresentado por estes autores para o cálculo de um estimador ML com uma distribuição exponencial generalizada para o erro de medição. O estudo considera um conjunto abrangente de distribuições para o ruído: distribuição Gaussiana, distribuição uniforme, distribuição t-Student, distribuição do qui-quadrado, distribuição F-Snedcor, distribuição Gama, e distribuição de Weibull. São usados 5% de *outliers* isolados gerados pela distribuição considerada com uma variância cinco vezes superior. Os resultados destacam, nos casos com *outliers* em que a distribuição do erro não é simétrica, o WARME seguido de perto pelo estimador ARMENSI. Embora nas restantes situações (com a excepção natural do caso da distribuição Gaussiana sem *outliers*, para o qual o estimador LS é óptimo) o estimador WARME apresente o melhor desempenho, a diferença é menos acentuada.

Não existe, tanto quanto se sabe, para o caso de resposta multivariada qualquer estudo de simulação que aflore o campo da estimação robusta de modelos não-lineares nos parâmetros.

3.2 Descrição das experiências

Como se viu na secção 1.2 na página 2 geram-se n observações, das quais n_c são observações "regulares", e $n_o = n - n_c$ podem representar *outliers*.[1] Os dados são criados de acordo com o modelo

$$\boldsymbol{y}_{i_c} = \boldsymbol{f}(\boldsymbol{x}_{i_c}, \boldsymbol{\theta}) + \boldsymbol{\epsilon}_{i_c}, \qquad i_c = 1, 2, \ldots, n_c, \tag{3.1}$$

e os *outliers* são gerados por

$$\boldsymbol{y}_{i_o} = \boldsymbol{f}(\boldsymbol{x}_{i_o}, \boldsymbol{\theta}) + \boldsymbol{\Delta}_{\boldsymbol{y}} + \boldsymbol{\epsilon}_{i_o}, \qquad i_o = 1, 2, \ldots, n_o, \tag{3.2}$$

onde os subíndices c e o designam, respectivamente, as observações regulares e os *outliers* presentes na amostra, $\boldsymbol{\Delta}_{\boldsymbol{y}}$ é uma perturbação na resposta, e as restantes variáveis têm o significado habitual.

No caso univariado a distância $\boldsymbol{\Delta}_{\boldsymbol{y}}$ é definida por $\delta_{\mathrm{R}}\sigma_\epsilon$, onde σ_ϵ designa o desvio padrão do erro de medida, ou seja, é o número de desvios padrão do erro que os *outliers* distam dos pontos regulares no espaço-y. No caso multivariado adopta-se

$$\boldsymbol{\Delta}_{\boldsymbol{y}} = \delta_{\mathrm{R}}[\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, \ldots, \sqrt{\sigma_{n_y n_y}}]^T,$$

onde δ_{R} é um escalar e σ_{jj} denota os elementos da diagonal principal da matriz de covariâncias $\boldsymbol{\Sigma}$.

A localização dos *outliers*, ou seja, a definição do conjunto de valores de i_o, foi efectuada de forma arbitrária.

[1] O delineamento das simulações descrito na presente secção é uma adaptação de aspectos do planeamento de experiências usado em Sebert *et al.* (1998) e Wisnowski *et al.* (2001).

Para conferir um maior cunho de verosimilhança aos resultados, as simulações aqui analisadas retratam casos reais apresentados na literatura. Como tal, os valores especificados para $\boldsymbol{\theta}$ e para o parâmetro de escala das várias distribuições do erro aleatório investigadas são estimados com base nos dados experimentais; as variáveis independentes $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ e a dimensão da amostra, n, tomam os valores originais.

As distribuições consideradas para gerar o erro de medição $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n$ foram:

Caso univariado

1. Gaussiana N(0, σ_ϵ).

2. Gaussiana contaminada CN(t, λ), em que $\lambda > 0$ que designa a distribuição N(0, σ_ϵ) contaminada pela distribuição N(0, $\lambda\sigma_\epsilon$) com probabilidade t; usamos, em particular, CN(0,10, 2) e CN(0,30, 5) para representar, respectivamente, situações de contaminação moderada e elevada.

3. Gaussiana enviesada (Azzalini, 1985, 1986), uma distribuição unimodal indexada por três parâmetros, o parâmetro de localização, o parâmetro de dispersão, e o parâmetro de forma (indicando o grau de assimetria da distribuição). Esta distribuição pode ser considerada uma generalização da distribuição Gaussiana a que se reduz quando o valor do parâmetro de forma é nulo. Existe uma parametrização alternativa — denominada parametrização centrada e denotada por SN(μ, σ, γ) — que integra a média, μ, o desvio padrão, σ, e o coeficiente de assimetria, γ. Neste trabalho usamos média nula, $\sigma = \sigma_\epsilon$, e $\gamma = 0{,}8$.

4. Cauchy (μ, σ) com parâmetro de localização, μ, nulo. Seguindo Gonin e Money (1989, p. 237) o parâmetro de escala σ é calculado de modo que o percentil 95 coincida com o quantil 97,5 da distribuição Gaussiana, ou seja, $\sigma = 0{,}31\sigma_\epsilon$.

Caso multivariado

1. Gaussiana multivariada $\mathrm{N}_{n_y}(\mathbf{0}, \boldsymbol{\Sigma})$.

2. Gaussiana multivariada $\mathrm{N}_{n_y}(\mathbf{0}, \boldsymbol{\Sigma})$ contaminada pela distribuição $\mathrm{N}_{n_y}(\mathbf{0}, \lambda\boldsymbol{\Sigma})$ com probabilidade t. Novamente duas situações foram estudadas; nível de contaminação moderado, isto é $t = 0{,}10$, $\lambda = 2$ e nível de contaminação elevado, nomeadamente $t = 0{,}30$, $\lambda = 5$.

3. t-Student multivariada $t_p(k; \boldsymbol{\lambda}, \boldsymbol{\Delta})$, onde $k > 0$ é o número de graus de liberdade, $\boldsymbol{\lambda} \in \mathbb{R}^p$ é o parâmetro de localização, e $\boldsymbol{\Delta}$ é uma matriz $p \times p$ simétrica definida positiva; aqui usou-se $t_{n_y}(3; \mathbf{0}, \boldsymbol{\Sigma}/3)$.

A distribuição Gaussiana contaminada é um clássico nos estudos empíricos de estimadores robustos; usa-se a distribuição Gaussiana enviesada para investigar o efeito de erros distribuídos de forma assimétrica, e a distribuição de Cauchy exemplifica comportamentos "patológicos". A figura 3.1 na próxima página ilustra as funções densidade das distribuições univariadas atrás indicadas.

A tabela 3.1 na página seguinte mostra os cenários referentes à combinação de diferentes níveis de percentagem de *outliers* e da distância normalizada δ_{R} considerados

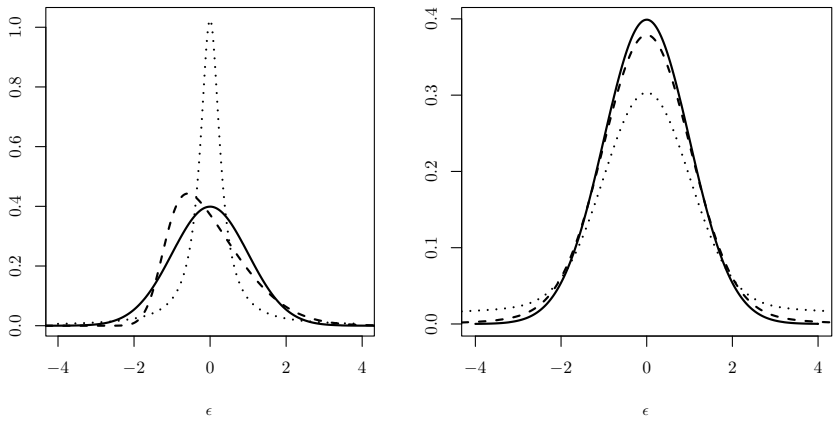


Figura 3.1 Funções densidade da distribuição Gaussiana estandardizada (a cheio); da distribuição Gaussiana enviesada com parametrização centrada SN(0, 1, 0,8) (esquerda, a tracejado); da distribuição de Cauchy (0, 0,31) (esquerda, a ponteado); da distribuição Gaussiana contaminada CN(0,10, 2) (direita, a tracejado); e de CN(0,30, 5) (direita, a ponteado).

Tabela 3.1 Cenários simulados

Factor	Nível				
Proporção de *outliers* (%)	0	10	20	15	30
δ_{R}		5	5	10	10

neste estudo. Em cada cenário foram geradas observações com as distribuições referidas anteriormente, e estudaram-se os seguintes estimadores:

Caso univariado

1. mínimos quadrados (LS),
2. norma L_p com $1 \leqslant p \leqslant 1{,}9$,
3. mediana dos quadrados mínima (LMS),
4. quadrados aparados mínimos com 25% de aparamento (LTS25),
5. quadrados aparados mínimos com 50% de aparamento (LTS50),
6. MM com 95% de eficiência iniciado a partir da estimativa LMS (MM-LMS),
7. MM com 95% de eficiência iniciado a partir da estimativa LTS25 (MM-LTS25),
8. MM com 95% de eficiência iniciado a partir da estimativa LTS50 (MM-LTS50),
9. τ,
10. diferenças aparadas mínimas com 25% de aparamento (LTD25),
11. diferenças aparadas mínimas com 50% de aparamento (LTD50).

Caso multivariado

1. critério do determinante (ML),
2. desvios absolutos mínimos multivariado (norma ℓ_1 de ℓ_2) (LAD),
3. M multivariado,
4. mediana da verosimilhança máxima (MML),
5. máxima verosimilhança aparada com 25% de aparamento (MTL25),
6. máxima verosimilhança aparada com 50% de aparamento (MTL50).

Por fim, referem-se vários aspectos computacionais.

- Cada simulação é composta por 100 réplicas, o limite prático imposto pela potência computacional da estação de trabalho com processador Pentium 4 a 1,8 GHz utilizada.
- A solução numérica de modelos compostos por um sistema de equações diferenciais ordinárias é obtida pela rotina de integração `LSODA` (Hindmarsh, 1983).
- Na utilização do algoritmo de optimização MDE estabeleceu-se 10^{-8} para valor da tolerância do critério de paragem, porquanto constitui um compromisso aceitável entre a qualidade dos resultados e o tempo de execução das simulações. Para cada um dos casos apresentados mais à frente, os restantes parâmetros algorítmicos são idênticos aos indicados na secção de aplicação aos dados experimentais respectiva.
- Tendo em conta o referido em 2.4.3 na página 33, é desejável incluir o "verdadeiro" valor de $\boldsymbol{\theta}$ na população inicial do algoritmo de optimização.
- No caso univariado as estimativas LS obtidas a partir dos dados experimentais[2] foram tomadas como os valores "verdadeiros" de $\boldsymbol{\theta}$, e, de acordo com o que acabou de afirmar-se, são adicionadas à população inicial do algoritmo MDE. No caso multivariado este papel foi atribuído às estimativas ML (critério do determinante).
- Para o desvio padrão do erro de medida, σ_ϵ, usou-se a estimativa LS dada por (2.3), enquanto que a matriz de covariâncias $\boldsymbol{\Sigma}$ foi estimada recorrendo a (2.26).
- O erro de medida Gaussiano e de Cauchy foi gerado pelas funções `rnorm` e `rcauchy`, respectivamente. O erro que segue a distribuição Gaussiana enviesada foi obtidos usando a função `rsn` da programateca `sn`. No caso multivariado, o erro de medida Gaussiano e t-Student foi gerado, respectivamente, pelas funções `rmvnorm` e `rmvt` da programateca mvtnorm.
- Na inicialização do gerador de números aleatórios estabeleceu-se 0 como semente.

[2]Em geral, as estimativas obtidas neste trabalho são basicamente idênticas às estimativas recolhidas na literatura.

3.3 Critérios de comparação

Sem dúvida que o critério mais vulgarmente usado para avaliar o desempenho de estimadores em estudos de simulação é o MSE, o qual combina o enviesamento do estimador, $\hat{\theta} - \theta$, com a sua variância $\sigma^2_{\hat{\theta}}$, visto que, como é bem conhecido,

$$\text{MSE} = E[(\hat{\theta} - \theta)^2] = (\hat{\theta} - \theta)^2 + \sigma^2_{\hat{\theta}}.$$

Porém, alguns autores preconizam o uso de critérios robustos. Assim, seguindo You (1999), vão-se usar medidas robustas análogas do MSE e do enviesamento. Para medida de enviesamento, uma alternativa é RB = $\text{med}(\hat{\theta}) - \theta$ e como análogo do MSE pode usar-se $\text{med}(|\hat{\theta} - \theta|)$. Concretamente, aqui, à semelhança do que é feito no trabalho de Midi (1999), vai-se utilizar uma medida de eficiência dos estimadores em relação ao estimador LS aplicado a dados apenas com erro Gaussiano, definida por

$$\frac{\text{med}(|\hat{\theta}_{\text{LS}} - \theta|)}{\text{med}(|\hat{\theta} - \theta|)}.$$

Para o caso dos modelos com resposta multivariada pode usar-se o estimador ML em vez do estimador LS como no caso univariado. O critério RB é normalizado pelo módulo de θ.

Além disso, é necessário avaliar o erro de amostragem destas duas estatísticas. Para o efeito, Lewis e Orav (1989, p. 275) sugerem a técnica do *bootstrap* em estudos de simulação em que razões de morosidade computacional implicam o uso de um número baixo de réplicas como é o caso deste trabalho. Deste modo, usa-se o intervalo de confiança a 95% calculado usando o método do percentil com 999 amostras na simulação *bootstrap*. Note-se que a amplitude do intervalo de confiança mede a precisão com que a eficiência e o enviesamento são determinados. Os cálculos necessários foram feitos com recurso às funções `boot` e `boot.ci` da programateca `boot` (Davison e Hinkley, 1997).

Capítulo 4

Aplicações da simulação de Monte Carlo a modelos com resposta univariada

4.1 Isomerização catalítica do n-pentano

4.1.1 Descrição genérica do problema

Carr (1960) estudou a isomerização catalítica do pentano na presença de hidrogénio. O catalisador é um metal depositado num suporte refractário. Os dados experimentais reproduzidos na tabela 4.1 na página seguinte foram obtidos num reactor diferencial, operado à temperatura de 750°F.

Uma das equações cinéticas propostas é

$$r' = \frac{kK_{\mathrm{C5}}(p_{\mathrm{C5}} - p_{\mathrm{iC5}}/K_p)}{1 + K_{\mathrm{H}}p_{\mathrm{H}} + K_{\mathrm{C5}}p_{\mathrm{C5}} + K_{\mathrm{iC5}}p_{\mathrm{iC5}}}, \tag{4.1}$$

onde os índices H, C5, e iC5 indicam o hidrogénio, n-pentano, e isopentano, respectivamente, k é uma constante que depende do catalisador e da temperatura, p representa a pressão parcial (psia), K representa a constante de adsorção de equilíbrio (psia^{-1}), K_p designa a constante de equilíbrio, na base da pressão, e r', expresso em $\mathrm{g_{pentano}}/(\mathrm{g_{cat.}}\ \mathrm{h})$, designa a razão da velocidade de reacção pela massa de catalisador. A constante de equilíbrio da reacção é conhecida sendo $K_p = 1{,}632$ a 750°F.

A análise deste caso feita por Bates e Watts (1988, pp. 55–58 e 211–213), nomeadamente a ausência de limite superior dos intervalos de verosimilhança e a presença de correlações elevadas entre as estimativas dos parâmetros, mostra a qualidade estatística insatisfatória do modelo acabado de descrever. O aumento da precisão das estimativas pode ser conseguido planeando experiências adicionais; uma outra possibilidade, conforme foi referido na página 7, consiste na reparametrização de (4.1), que pode ser reescrita de forma a eliminar-se o produto kK_{C5}, obtendo-se

$$r' = \frac{p_{\mathrm{C5}} - p_{\mathrm{iC5}}/1{,}632}{\beta_1 + \beta_2 p_{\mathrm{H}} + \beta_3 p_{\mathrm{C5}} + \beta_4 p_{\mathrm{iC5}}}, \tag{4.2}$$

onde

$$k = 1/\beta_3, \qquad K_{\mathrm{H}} = \beta_2/\beta_1, \qquad K_{\mathrm{C5}} = \beta_3/\beta_1, \qquad K_{\mathrm{iC5}} = \beta_4/\beta_1.$$

A formulação que acabou de descrever-se apresenta uma qualidade estatística satisfatória no que diz respeito aos parâmetros β_2, β_3 e β_4. Porém, o intervalo de confiança de β_1 inclui valores negativos, sem significado físico (Bates e Watts, 1988, pp. 213 e 214). Por isso, estes autores chamam a atenção para o "valor limitado" dos dados experimentais.

Tabela 4.1 Isomerização do n-pentano: dados experimentais

	pressão parcial/psia			
Obs.	hidrogénio	n-pentano	isopentano	velocidade/h^{-1}
1	205,8	90,9	37,1	3,541
2	404,8	92,9	36,3	2,397
3	209,7	174,9	49,4	6,694
4	401,6	187,2	44,9	4,722
5	224,9	92,7	116,3	0,593
6	402,6	102,2	128,9	0,268
7	212,7	186,9	134,4	2,797
8	406,2	192,6	134,9	2,451
9	133,3	140,8	87,6	3,196
10	470,9	144,2	86,9	2,021
11	300,0	68,3	81,7	0,896
12	301,6	214,6	101,7	5,084
13	297,3	142,2	10,5	5,686
14	314,0	146,7	157,1	1,193
15	305,7	142,0	86,0	2,648
16	300,1	143,7	90,2	3,303
17	305,4	141,1	87,4	3,054
18	305,2	141,5	87,0	3,302
19	300,1	83,0	66,4	1,271
20	106,6	209,6	33,0	11,648
21	417,2	83,9	32,9	2,002
22	251,0	294,4	41,5	9,604
23	250,3	148,0	14,7	7,754
24	145,1	291,0	50,2	11,590

Tabela 4.2 Isomerização do n-pentano: estimativas para os parâmetros do modelo

	Parâmetros			
Estimadores	β_1/h psia	β_2/h	β_3/h	β_4/h
LS	0,73786	0,052273	0,027839	0,12332
L_p	0,72432	0,052562	0,027691	0,12354
LMS	0,49468	0,05758	0,023581	0,13901
LTS25	0,9385	0,059326	0,023882	0,12068
LTS50	$6{,}8893\times10^{-14}$	0,0564	0,026128	0,1455
MM-LTS25	0,62938	0,054821	0,027662	0,11878
τ	0,57097	0,054277	0,027237	0,12472
LTD25	0,76418	0,050141	0,040812	0,027871
LTD50	0,77446	0,054234	0,033289	0,067774

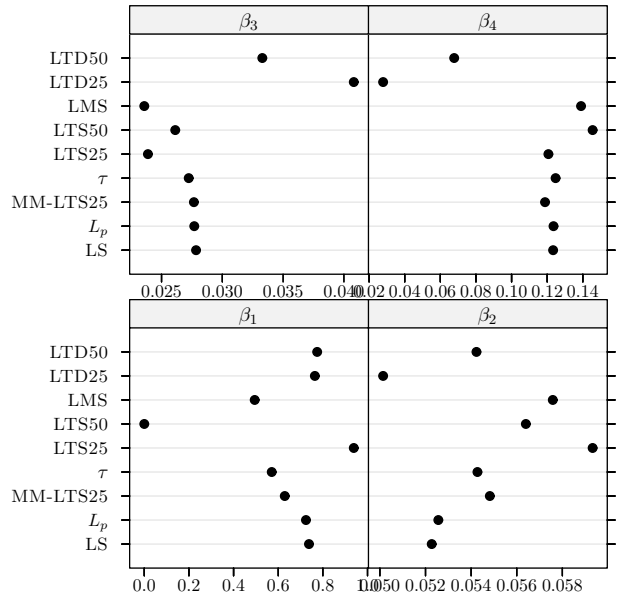


Figura 4.1 Isomerização do n-pentano: estimativas para os parâmetros do modelo.

4.1.2 Aplicação aos dados experimentais

Na tabela 4.2 encontram-se as estimativas dos parâmetros do modelo (4.2) de oito estimadores robustos mais o estimador LS aplicados às 24 observações experimentais. O valor obtido para a estimativa do desvio padrão do erro de medição é de 0,4021 h^{-1}. Na figura 4.1 apresentam-se gráficos para os resultados referentes à tabela 4.2.

Os valores usados para os parâmetros do algoritmo MDE são os valores por omissão referidos na respectiva descrição na página 34, designadamente: $N_\mathrm{P} = 41$, $F_\mathrm{b} = 0{,}5$, $C_\mathrm{R} = 0{,}8$, e $\xi = 1{,}5$. A tolerância usada para o critério de paragem foi 10^{-15}, e o espaço de procura é $[10^{-80}, 10^{-80}, 10^{-80}, 10^{-80}] \leqslant \boldsymbol{\beta}^T \leqslant [100, 10, 10, 10]$[1]. Finalmente,

[1] Na prática substituiu-se zero por um valor positivo muito pequeno, por forma a garantir que o valor da função objectivo no passo 3 do estimador MM seja finito (limitação imposta pela rotina de optimização L-BFGS-B).

foi incluída na população inicial a estimativa $\hat{\boldsymbol{\beta}} = [0{,}73738, 0{,}052274, 0{,}027841, 0{,}12331]^T$ apresentada por Huet *et al.* (1996, p. 47, tabela 2.5) em conformidade com o exposto em 2.4.3 na página 33.

Vê-se que as estimativas LS, L_p, MM-LTS25, e τ são basicamente idênticas. Por outro lado, na maioria dos casos as estimativas LMS, LTS, e LTD apresentam diferenças nítidas das restantes, salientando-se as estimativas LTD de β_4, e LMS e LTS50 de β_1, as quais se afastam acentuadamente. Neste último caso, a razão prende-se, possivelmente, com a conjugação da fraca qualidade estatística de β_1 ao modo de definição dos estimadores LMS e LTS50, que, grosso modo, "ignora" metade da amostra. Para finalizar, importa notar que uma análise em termos mais rigorosos só pode efectuar-se quando forem desenvolvidos métodos fiáveis de determinação de intervalos de confiança para regressão não-linear robusta.

4.2 Oxidação catalítica do propeno

4.2.1 Descrição genérica do problema

Na tabela 4.3 na próxima página encontram-se os dados da oxidação catalítica do propeno obtidos às temperaturas de 350 °C, 375 °C, e 390 °C (Watts, 1994). O catalisador é molibdato de bismuto num suporte de sílica.

Tan *et al.* (1988) desenvolveram e analisaram um conjunto de sete modelos cinéticos candidatos. A lei de velocidade seleccionada pelos autores é

$$-r_{\mathrm{C_3H_6}} = \frac{k_\mathrm{a} k_\mathrm{r} c_{\mathrm{O_2}}^{0,5} c_{\mathrm{C_3H_6}}}{k_\mathrm{a} c_{\mathrm{O_2}}^{0,5} + n k_\mathrm{r} c_{\mathrm{C_3H_6}}}, \tag{4.3}$$

onde c representa concentração (mmol/dm^3), $r_{\mathrm{C_3H_6}}$ representa a velocidade de consumo do propeno (mmol/kg s), k_a ((mmol dm^3)^{0,5}/kg s) e k_r (dm^3/kg s) designam as constantes de velocidade para a adsorção de oxigénio e para a oxidação do propeno, respectivamente, e, por último, n é o número estequiométrico, isto é, o número de moles de oxigénio consumidas por mole de propeno oxidado.

A dependência entre as constantes de velocidade e a temperatura é descrita pela lei de Arrhenius

$$k = A \exp\Big(-\frac{E}{RT}\Big), \tag{4.4}$$

onde A é o factor de frequência ou pré-exponencial, E a energia de activação, R a constante dos gases, e T a temperatura absoluta. Um problema com esta formulação é que, de um modo geral, a estimação do factor de frequência e da energia de activação resulta em estimativas fortemente correlacionadas. O método usado para ultrapassar esta dificuldade é a reparametrização da expressão (4.4). Além disso, o recurso à transformação da forma "canónica" da lei de Arrhenius conduz a parâmetros com um comportamento estatístico superior ao dos parâmetros originais (Watts, 1994). No presente trabalho utiliza-se a seguinte reparametrização sugerida por Lohmann *et al.* (1992):

$$\alpha = \ln A - \frac{E}{RT_1}, \quad \beta = \ln A - \frac{E}{RT_2}, \tag{4.5}$$

Tabela 4.3 Oxidação do propeno: dependência entre a velocidade de consumo do propeno (mmol/kg s) e a concentração de propeno, $c_{C_3H_6}$, (mmol/dm^3) a concentração de oxigénio, c_{O_2}, (mmol/dm^3), e o número estequiométrico, n, a 350 °C, 375 °C, e 390 °C

$T = 350$ °C					$T = 375$ °C					$T = 390$ °C				
Obs.	$c_{C_3H_6}$	c_{O_2}	n	$-r_{C_3H_6}$	Obs.	$c_{C_3H_6}$	c_{O_2}	n	$-r_{C_3H_6}$	Obs.	$c_{C_3H_6}$	c_{O_2}	n	$-r_{C_3H_6}$
1	3,05	3,07	2,73	0,658	24	2,94	2,96	2,37	1,160	46	2,62	3,66	1,95	1,480
2	1,37	3,18	2,86	0,439	25	1,35	3,06	2,58	0,680	47	2,79	2,96	2,00	1,510
3	3,17	1,24	3,00	0,452	26	3,04	1,19	2,24	0,740	48	3,02	6,12	1,92	1,800
4	3,02	3,85	2,64	0,695	27	2,90	3,70	2,19	1,170	49	3,07	7,32	1,96	1,900
5	4,31	3,15	2,60	0,635	28	4,14	3,03	2,32	1,390	50	1,36	7,52	2,36	0,990
6	2,78	3,89	2,73	0,670	29	2,69	3,76	2,31	1,190	51	1,31	1,12	2,33	0,805
7	3,11	6,48	2,56	0,760	30	2,99	6,23	2,16	1,290	52	1,42	7,47	2,26	0,991
8	2,96	3,13	2,69	0,642	31	2,85	3,03	2,25	1,130	53	2,72	3,48	1,93	1,520
9	2,84	3,14	2,77	0,665	32	5,46	7,46	1,93	2,030	54	6,86	2,86	2,06	2,210
10	1,46	7,93	2,91	0,525	33	1,39	7,67	2,63	0,804	55	7,13	2,89	2,10	2,300
11	1,38	7,79	2,87	0,483	34	1,34	1,15	2,58	0,630	56	1,32	7,48	2,36	0,936
12	1,42	8,03	2,97	0,522	35	2,73	3,02	2,16	1,080	57	7,09	3,27	2,16	2,430
13	1,49	7,78	2,93	0,530	36	1,46	7,65	2,64	0,864	58	2,88	3,76	1,85	1,640
14	3,01	3,03	2,75	0,635	37	1,39	7,56	2,53	0,772	59	1,33	7,84	2,38	0,975
15	1,35	8,00	2,90	0,480	38	1,33	7,49	2,64	0,777	60	7,14	3,22	2,10	2,300
16	1,52	8,22	2,94	0,544	39	1,37	7,75	2,51	0,745	61	1,37	7,89	2,39	0,996
17	5,95	6,13	2,38	0,893	40	7,02	2,93	2,25	1,310	62	5,39	7,25	1,76	2,760
18	1,46	8,41	2,89	0,517	41	2,89	2,91	2,27	1,160	63	1,31	2,90	2,28	0,823
19	5,68	7,75	2,41	0,996	42	7,30	2,96	2,22	1,360	64	2,74	3,54	1,84	1,530
20	1,36	3,10	2,81	0,416	43	1,35	7,66	2,55	0,741	65	2,89	7,48	1,83	1,790
21	1,42	1,25	2,86	0,367	44	3,15	7,52	2,14	1,440	66	5,29	7,23	1,75	2,760
22	3,18	7,89	2,59	0,835	45	2,75	2,93	2,15	1,050					
23	2,87	3,06	2,76	0,609										

Fonte: Apêndice 1 de Watts (1994, p. 709). Note-se que no original as colunas n e $-r_{C_3H_6}$ encontram-se trocadas.

Tabela 4.4 Oxidação do propeno: parâmetros do algoritmo de optimização MDE usados na regressão dos dados experimentais

Espaço de procura $\ln([0{,}01, 0{,}01, 0{,}001, 0{,}001]) \leqslant [\alpha_\mathrm{a}, \beta_\mathrm{a}, \alpha_\mathrm{r}, \beta_\mathrm{r}] \leqslant \ln([100, 100, 10, 10])$					
N_P	F_b	C_R	ξ	Tolerância	Ponto incluído na população inicial
41	0,5	0,8	1,5	10^{-15}	$[0{,}2263, 1{,}4558, -0{,}4387, -0{,}1115]$[a]

[a] Os valores de α_a, β_a, α_r, e β_r foram calculados a partir da parametrização $\ln k = \ln A^* - \frac{E}{R}(\frac{1}{T} - \frac{1}{T_o})$, onde $T_o = 648\,\mathrm{K}$, com as seguintes estimativas (Watts, 1994): $A^*_\mathrm{a} = 2{,}74\ (\mathrm{mmol\,dm^3})^{0,5}/\mathrm{kg\,s}$, $E_\mathrm{a} = 105{,}6\,\mathrm{kJ/mol}$, $A^*_\mathrm{r} = 0{,}794\,\mathrm{dm^3/kg\,s}$, e $E_\mathrm{r} = 28{,}1\,\mathrm{kJ/mol}$. Note-se que as unidades originais apresentam erros na ordem de grandeza que foram aqui eliminados.

onde T_1 e T_2 são temperaturas de referência, e α e β designam o logaritmo de uma constante de velocidade para duas temperaturas diferentes. Na maioria dos casos T_1 e T_2 tomam os valores extremos da gama de temperatura correspondente aos dados experimentais — neste caso $T_1 = 350\,^\circ\mathrm{C}$ e $T_2 = 390\,^\circ\mathrm{C}$. A constante de velocidade k para a temperatura T é definida por

$$k = \exp(\alpha\tau(T) + \beta(1 - \tau(T)) \tag{4.6}$$

com

$$\tau(T) = \frac{T_1}{T}\frac{(T - T_2)}{(T_1 - T_2)}. \tag{4.7}$$

Nas expressões anteriores a temperatura é expressa em kelvin.

4.2.2 Aplicação aos dados experimentais

A tabela 4.5 na página ao lado apresenta as estimativas dos parâmetros do modelo (4.3) obtidas a partir dos dados experimentais, com os valores dos parâmetros do algoritmo MDE que se encontram na tabela 4.4. O valor obtido para a estimativa do desvio padrão do erro de medição é de 0,06545 mmol $\mathrm{kg^{-1}\,s^{-1}}$. Na figura 4.2 na página ao lado apresentam-se gráficos para os resultados referentes à tabela 4.5 na próxima página.

Relativamente a α_r e β_r os resultados mostram que as estimativas LTD se afastam significativamente das outras. As diferenças entre as estimativas LS, L_p, LMS, LTS, MM-LTS25, e τ de β_a e β_r, são basicamente insignificantes. Contudo, pode constatar-se que existe para α_a e α_r um desvio nítido entre as estimativas de mínimos quadrados e as robustas por uma parte, e entre os resultados dos estimadores L_p e MM-LTS25 e os dos estimadores LMS, LTS, τ, por outra parte.

Tabela 4.5 Oxidação do propeno: estimativas para os parâmetros do modelo

	Parâmetros			
Estimadores	α_a	β_a	α_r	β_r
LS	0,2322	1,446	−0,4625	−0,1122
L_p	0,326	1,435	−0,5259	−0,1077
LMS	0,5371	1,465	−0,6675	−0,1413
LTS25	0,5224	1,438	−0,6817	−0,1142
LTS50	0,535	1,526	−0,6624	−0,165
MM-LTS25	0,4083	1,479	−0,5985	−0,1323
τ	0,5247	1,444	−0,6653	−0,1142
LTD25	0,4526	1,593	−0,8752	−0,2565
LTD50	0,4471	1,491	−0,7871	−0,1908

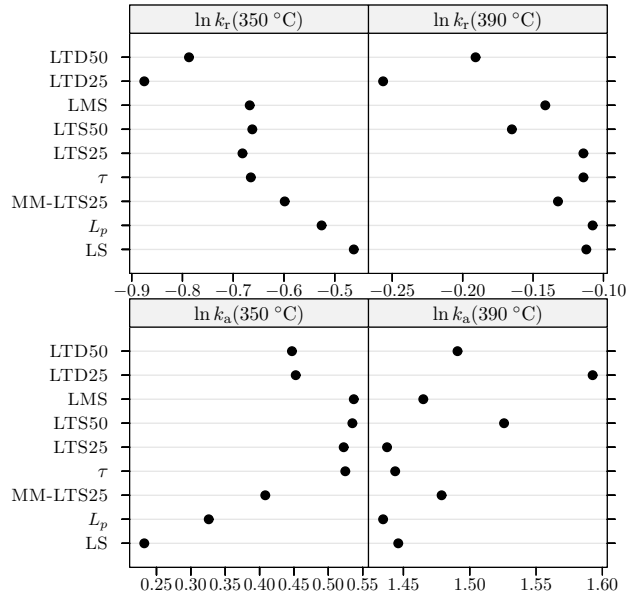


Figura 4.2 Oxidação do propeno: estimativas para os parâmetros do modelo.

Tabela 4.6 Dados de equilíbrio líquido-vapor do sistema $CH_3OCH_3(1)$ + $CF_3CH_2CF_3(2)$ a 303,68 K

Obs.	p/kPa	x	y
1	419,5	0,5315	0,6664
2	488,1	0,6743	0,8184
3	540,9	0,7624	0,8877
4	620,7	0,8859	0,9588
5	329,6	0,0829	0,0970
6	336,4	0,1708	0,2016
7	364,9	0,3558	0,4428
8	392,5	0,4553	0,5774
9	441,8	0,5820	0,7298
10	521,8	0,7305	0,8656
11	598,6	0,8517	0,9426

4.3 Regressão de dados de equilíbrio líquido-vapor da mistura refrigerante RE170 + R236fa

4.3.1 Descrição genérica do problema

No âmbito dos esforços que têm vindo a ser desenvolvidos para a substituição de misturas refrigerantes bem conhecidas por alternativas ambientalmente mais seguras, Bobbo *et al.* (1998) estudaram a mistura éter dimetílico (RE170) + 1,1,1,3,3,3-hexafluoropropano (R236fa). O quadro 4.6 fornece dados isotérmicos referentes ao equilíbrio líquido-vapor a 303,68 K.

O ajuste dos dados experimentais usa a equação de estado de Carnahan-Starling-De Santis (De Santis *et al.*, 1976), que pode ser escrita da seguinte forma:

$$p = \frac{RT}{V}\frac{1+Y+Y^2+Y^3}{(1-Y)^3} - \frac{a}{V(V+b)}, \quad \text{com} \quad Y = \frac{b}{4V}, \tag{4.8}$$

onde p designa a pressão, T a temperatura, V o volume molar, R a constante dos gases, e as constantes a e b são específicas para cada substância. Quando a equação de estado é aplicada a uma mistura, as constantes a e b são obtidas a partir de regras de mistura, neste caso, as regras de van der Waals:

$$a = z_1^2 a_1 + 2z_1 z_2 \sqrt{a_1 a_2}(1 - k_{12}) + z_2^2 a_2, \tag{4.9a}$$

$$b = z_1 b_1 + z_2 b_2, \tag{4.9b}$$

onde os parâmetros a_k e b_k se referem aos componentes puros, z_k representa a fracção molar quer da fase líquida quer da fase de vapor, e k_{12} é o parâmetro empírico de interacção binária.

A dependência da temperatura das constantes a e b é correlacionada pelas seguintes expressões empíricas:

$$a(T) = a_0 \exp(a_1 T + a_2 T^2), \tag{4.10a}$$

$$b(T) = b_0 + b_1 T + b_2 T^2. \tag{4.10b}$$

Tabela 4.7 Parâmetros para as equações (4.10a) e (4.10b) usados na regressão dos dados. Dados extraídos da tabela 2 de Bobbo *et al.* (1998, p. 1045)[a]

	Composto	
Parâmetro	CH_3OCH_3	$CF_3CH_2CF_3$
a_0/kPa dm^{6}mol^{-2}	$3{,}20884 \times 10^3$	$5{,}812833 \times 10^3$
a_1/K^{-1}	$-3{,}20482 \times 10^{-3}$	$-2{,}860835 \times 10^{-3}$
a_2/K^{-2}	$1{,}50810 \times 10^{-7}$	$-1{,}409685 \times 10^{-6}$
b_0/dm^3 mol^{-1}	$1{,}30775 \times 10^{-1}$	$1{,}976126 \times 10^{-1}$
b_1/dm^3 mol^{-1}K^{-1}	$-1{,}97630 \times 10^{-4}$	$-1{,}906306 \times 10^{-4}$
b_2/dm^3 mol^{-1}K^{-2}	$5{,}4664 \times 10^{-8}$	$-1{,}462412 \times 10^{-7}$

[a] Note-se que no original as colunas de valores encontram-se trocadas.

A tabela 4.7 apresenta os coeficientes correspondentes.

Repare-se que a equação (4.8) pode ser reescrita de forma a obter-se um polinómio de grau 5 no volume, como se segue:

$$RTb^4 - ab^3 + (12ab^2 - b^4p - 3RTb^3)V + (11b^3p - 48ab - 20RTb^2)V^2 + (64a - 80RTb - 36b^2p)V^3 + (16bp - 64RT)V^4 + 64pV^5 = 0, \quad (4.11)$$

com $V \neq b/4$.

No trabalho de Bobbo *et al.* (1998) a estimativa do parâmetro de interacção binária, k_{12}, define-se pela minimização da função objectivo

$$\sum_{i=1}^{n} \left(\frac{p_{\mathrm{exp},i} - p_{\mathrm{calc},i}}{p_{\mathrm{exp},i}} \right)^2, \quad (4.12)$$

onde os índices calc e exp designam valores calculados e experimentais, respectivamente. Assim, à luz da expressão anterior vai usar-se

$$\epsilon_i = (p_{\mathrm{exp},i} - p_{\mathrm{calc},i})/p_{\mathrm{exp},i} \quad (4.13)$$

na definição dos vários estimadores examinados neste capítulo.

No cálculo dos resíduos p_{calc} obtém-se a partir de cálculos de *ponto de bolha (pressão)*, o que dada a sua natureza iterativa implica um esforço computacional significativo. O coeficiente de partição líquido-vapor, K_k, é dado por

$$K_k = y_k/x_k = \exp(\ln \phi_k^L - \ln \phi_k^V) \quad (4.14)$$

com (De Santis *et al.*, 1976),

$$\ln \phi_k = \frac{4Y - 3Y^2}{(1-Y)^2} + \frac{b_k}{b}\frac{4Y - 2Y^2}{(1-Y)^3} + \frac{2}{RTb}\sum_{m=1}^{2} z_m a_{km} \ln \frac{V}{V+b} + \frac{b_k a}{RTb^2} \ln \frac{V+b}{V} - \frac{1}{RT}\frac{b_k a}{bV + b^2} - \ln \frac{pV}{RT}, \quad (4.15)$$

onde os índices L e V designam a fase líquida e de vapor, respectivamente, ϕ_k é o coeficiente de fugacidade do componente k, e x_k e y_k representam a fracção molar do componente k na fase líquida e de vapor, respectivamente. O procedimento de cálculo do coeficiente de fugacidade de cada componente na fase líquida ou de vapor para valores específicos de T, p, e z_1 e z_2, consiste nos seguintes passos:

1. Obter os parâmetros a_k e b_k correspondentes aos componentes puros usando as equações (4.9a) e (4.9b).

2. Calcular os parâmetros a e b para a mistura a partir das equações (4.10a) e (4.10b) com $z_k = x_k$ para a fase líquida e $z_k = y_k$ para a fase de vapor.

3. Determinar as raízes reais da equação (4.11); V^L corresponde à raiz de maior valor, V^V à raiz de menor valor.

4. Calcular o coeficiente de fugacidade usando a equação (4.15) para cada substância com os valores obtidos nos itens anteriores.

Aspectos computacionais. A implementação do algoritmo de ponto de bolha (pressão) adapta o programa VLMU discutido em Sandler (1999, apêndice A7.2, p. 467). Os valores experimentais de pressão e fracção molar na fase de vapor são usados para estimativas iniciais do procedimento.

As raízes da forma (4.11) da equação de estado de Carnahan-Starling-De Santis são determinadas calculando os valores próprios da matriz

$$\begin{bmatrix} -a_0/a_4 & -a_1/a_4 & -a_2/a_4 & -a_3/a_4 & -1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

onde $a_0, \ldots, a_4$ representam os coeficientes dos termos (por ordem crescente dos graus) desta equação polinomial. A função `solve.polynomial` da programateca `polynom` implementa este método.

Quando é obtida uma só raiz, usa-se a pseudo-raiz duma fase artificial no cálculo dos coeficientes de partição, conforme o método proposto por Jovanović e Paunović (1984).

4.3.2 Aplicação aos dados experimentais

Na tabela 4.8 na página ao lado encontram-se os parâmetros do método MDE usados no cálculo das diversas estimativas de k_{12} a partir dos dados experimentais de equilíbrio líquido-vapor, obtendo-se os valores apresentados na tabela 4.9 na próxima página. O valor obtido para a estimativa do desvio padrão do erro de medição é de 0,007814.

É de destacar que os valores estimados pelo método LTD encontram-se notoriamente afastados dos restantes, que são praticamente idênticos.

Tabela 4.8 Equilíbrio líquido-vapor do sistema $CH_3OCH_3(1)$ + $CF_3CH_2CF_3(2)$: parâmetros do algoritmo de optimização MDE usados na regressão dos dados experimentais

f_0				Espaço de procura[a]	
$\sum_{i=1}^{n}\left(\frac{p_{\mathrm{exp},i}-\bar{p}_{\mathrm{exp}}}{\bar{p}_{\mathrm{exp}}}\right)^2$				$-3 \leqslant k_{12} \leqslant 0{,}3$	
N_P	F_b	C_R	ξ	Tolerância	Ponto incluído na população inicial
11	0,5	0,8	1,5	10^{-15}	$-0{,}0900$[b]

[a] Corresponde à região de convergência do algoritmo de ponto de bolha (pressão).
[b] Tabela 3 de Bobbo *et al.* (1998, p. 1046).

Tabela 4.9 Equilíbrio líquido-vapor do sistema $CH_3OCH_3(1)$ + $CF_3CH_2CF_3(2)$: estimativas de k_{12}

	Estimadores								
	LS	L_p	LMS	LTS25	LTS50	MM-LTS25	τ	LTD25	LTD50
k_{12}	$-0{,}09$	$-0{,}09$	$-0{,}091$	$-0{,}091$	$-0{,}092$	$-0{,}09$	$-0{,}09$	$-0{,}096$	$-0{,}1$

4.4 Regressão de dados de equilíbrio sólido-líquido do sistema binário 1,4-butanodiol + 4-metoxifenol

4.4.1 Descrição genérica do problema

Lee *et al.* (2001) obtiveram dados do equilíbrio sólido-líquido para o sistema binário 1,4-butanodiol + 4-metoxifenol apresentados na tabela 4.10 na página seguinte. Como se pode observar na figura 4.3 na próxima página onde se representa o diagrama de fases, a mistura é um sistema eutéctico simples.

O modelo considerado para a solubilidade de um sólido num líquido é descrito pela seguinte equação:

$$\ln(x_k\gamma_k) = \frac{\Delta H_k^{\mathrm{fus}}}{R}\Big(\frac{1}{T_{m,k}} - \frac{1}{T}\Big), \tag{4.16}$$

onde x_k designa a solubilidade do composto k sólido à temperatura T, R é a constante dos gases, e γ_k, $\Delta H_k^{\mathrm{fus}}$, e $T_{m,k}$ designam o coeficiente de actividade, a entalpia molar de fusão, e temperatura de fusão para o componente k, respectivamente.

O modelo de cálculo dos coeficientes de actividade é uma extensão da teoria das soluções regulares com parâmetros ajustáveis a misturas líquidas com componentes polares como o 1,4-butanodiol e o 4-metoxifenol. Precisando, para uma mistura binária tem-se:

$$\begin{aligned} RT\ln\gamma_1 &= V_1^L[(\delta_1-\delta_2)^2 + 2\lambda_{12}\delta_1\delta_2]\Phi_2^2, \\ RT\ln\gamma_2 &= V_2^L[(\delta_1-\delta_2)^2 + 2\lambda_{12}\delta_1\delta_2]\Phi_1^2, \end{aligned} \tag{4.17}$$

Tabela 4.10 Dados de equilíbrio sólido-líquido do sistema 1,4-butanodiol(1) + 4-metoxifenol(2)

Obs.	x_1	T/K
1	0,1597	317,8
2	0,2529	311,1
3	0,3246	304,9
4	0,3642	300,8
5	0,4585	290,5
6	0,4858	286,8
7	0,5032	285,0
8	0,5490	279,3
9	0,5975	271,3
10	0,6183	270,2
11	0,6408	272,7
12	0,6732	274,5
13	0,8008	282,7
14	0,8949	288,6

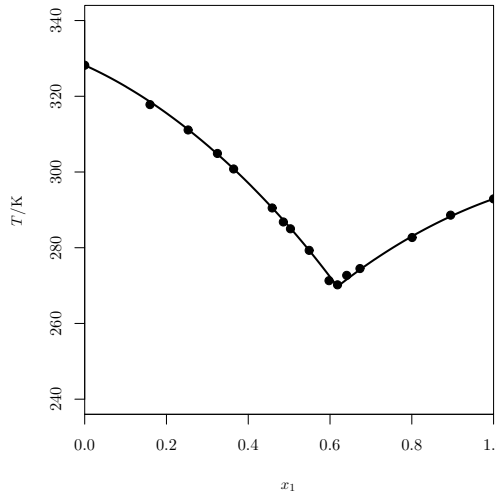


Figura 4.3 Diagrama de fases da mistura 1,4-butanodiol(1) + 4-metoxifenol(2). Os pontos representam os dados experimentais, enquanto que as curvas foram calculadas com $\lambda_{12} = -0{,}033$ segundo as equações (4.18a) e (4.18b).

Tabela 4.11 Propriedades das substâncias puras

Composto	T_m/K	ΔH^{fus}/kJ mol^{-1}	V^L/cm^3 mol^{-1} a 298,15 K	δ/(kJ cm^{-3})$^{1/2}$
1,4-butanodiol	292,9	18,70	88,974	0,784
4-metoxifenol	328,2	18,30	84,47	0,901

Fonte: Tabela 1 de Lee *et al.* (2001, p. 4597).

com

$$\Phi_1 = \frac{x_1 V_1^L}{x_1 V_1^L + x_2 V_2^L},$$
$$\Phi_2 = \frac{x_2 V_2^L}{x_1 V_1^L + x_2 V_2^L},$$

onde δ_k, Φ_k e V^L representam o parâmetro de solubilidade, a fracção volúmica e o volume molar como líquido do componente k, respectivamente, e λ_{12} é o parâmetro de interacção binária. Os volumes molares, as entalpias molares de fusão, as temperaturas de fusão, e os parâmetros de solubilidade das substâncias puras encontram-se na tabela 4.11.

Num sistema binário eutéctico simples podem distinguir-se três partes: uma de $(0, T_{m,2})$ ao ponto eutéctico $(x_{\text{eut}}, T_{\text{eut}})$, que se designa por zona II, outra do ponto eutéctico a $(1, T_{m,1})$, que se designa por zona I, e a terceira correspondente ao ponto eutéctico. Em termos mais precisos:

- Na zona II, que corresponde ao equilíbrio entre o sólido puro do componente 2 e o líquido, combinando as equações (4.17) e (4.16) para o componente 2, tem-se

$$T = \frac{V_2^L[(\delta_1 - \delta_2)^2 + 2\lambda_{12}\delta_1\delta_2]\Phi_1^2 + \Delta H_2^{\text{fus}}}{\Delta H_2^{\text{fus}}/T_{m,2} - R\ln(1 - x_1)}, \tag{4.18a}$$

  expressão que traduz a linha *liquidus* nesta região.

- Seguindo a mesma lógica, a linha *liquidus* na zona I, onde o equilíbrio se estabelece entre o sólido puro do componente 1 e o líquido, é dada por

$$T = \frac{V_1^L[(\delta_1 - \delta_2)^2 + 2\lambda_{12}\delta_1\delta_2]\Phi_2^2 + \Delta H_1^{\text{fus}}}{\Delta H_1^{\text{fus}}/T_{m,1} - R\ln x_1}. \tag{4.18b}$$

- O ponto eutéctico é definido pela intersecção das linhas *liquidus*; isto é a solução que satisfaz simultaneamente as duas relações antecedentes.

Note-se que o modelo termodinâmico, tal como acabou de descrever-se, tem uma estrutura não-linear no parâmetro λ_{12} unicamente no ponto eutéctico.

Lee *et al.* sugerem uma estimativa de λ_{12} que minimiza

$$\sum_{0<x_1<x_{\text{eut}}} \frac{|T_{\text{exp}} - T_{\text{calc}}^{\text{II}}|}{T_{\text{exp}}} + \sum_{x_{\text{eut}}<x_1<1} \frac{|T_{\text{exp}} - T_{\text{calc}}^{\text{I}}|}{T_{\text{exp}}}, \tag{4.19}$$

onde $T_{\text{calc}}^{\text{I}}$ e $T_{\text{calc}}^{\text{II}}$ são obtidos a partir das equações (4.18b) e (4.18a), respectivamente; x_{eut} pode calcular-se a partir da expressão que se obtém igualando o segundo membro das mesmas equações.

Uma formulação alternativa usada neste trabalho consiste em atribuir a x_{eut} o papel de variável de decisão, o que não obriga à resolução duma função não-linear para estimar o valor da composição do ponto eutéctico. De acordo com os dados experimentais, o valor de x_{eut} confina-se ao intervalo entre as composições dos pontos 9 e 11. Assim, em termos práticos, x_{eut} é basicamente uma variável de decisão dicotómica, a qual determina a localização do ponto 10 na zona I ou II de equilíbrio sólido-líquido.

Finalmente, nota-se que, de acordo com (4.19), se tem

$$\epsilon_i = \begin{cases} (T_{\text{exp},i} - T_{\text{calc},i}^{\text{II}})/T_{\text{exp},i} & \text{se } 0 < x_{1,i} < x_{\text{eut}} \\ (T_{\text{exp},i} - T_{\text{calc},i}^{\text{I}})/T_{\text{exp},i} & \text{se } x_{\text{eut}} < x_{1,i} < 1. \end{cases} \tag{4.20}$$

4.4.2 Aplicação aos dados experimentais

Na tabela 4.12 na próxima página encontram-se os parâmetros do método MDE usados no cálculo das várias estimativas de λ_{12} a partir dos dados experimentais de equilíbrio sólido-líquido. Os resultados apresentam-se na tabela 4.13 na página seguinte. O valor obtido para a estimativa do desvio padrão do erro de medição é de 0,001968.

Como se pode observar, os valores produzidos pelos nove métodos são essencialmente idênticos.

Tabela 4.12 Equilíbrio sólido-líquido do sistema 1,4-butanodiol(1) + 4-metoxifenol(2): parâmetros do algoritmo de optimização MDE usados na regressão dos dados experimentais

f_0				Espaço de procura	
$\sum_{i=1}^{n}\left(\frac{T_{\mathrm{exp},i}-\bar{T}_{\mathrm{exp}}}{\bar{T}_{\mathrm{exp}}}\right)^2$				$[-10, 0{,}5975] \leqslant [\lambda_{12}, x_{\mathrm{eut}}] \leqslant [10, 0{,}6408]$	
N_{P}	F_{b}	C_{R}	ξ	Tolerância	Ponto incluído na população inicial
21	0,5	0,8	1,5	10^{-15}	$[-0{,}033^{\mathrm{a}}, 0{,}6144^{\mathrm{b}}]$

[a] Tabela 5 de Lee *et al.* (2001, p. 4600).
[b] Tabela 3 de Lee *et al.* (2001, p. 4598).

Tabela 4.13 Equilíbrio sólido-líquido do sistema 1,4-butanodiol(1) + 4-metoxifenol(2): estimativas de λ_{12}

	Estimadores								
	LS	L_p	LMS	LTS25	LTS50	MM-LTS25	τ	LTD25	LTD50
λ_{12}	−0,033	−0,033	−0,033	−0,033	−0,033	−0,033	−0,033	−0,033	−0,034

4.5 Lixiviação de minério manganífero

4.5.1 Descrição genérica do problema

Vegliò *et al.* (2001a) e Vegliò *et al.* (2001b) propuseram o seguinte modelo cinético simplificado para a dissolução do dióxido de manganésio no tratamento por lixiviação de minérios de manganésio usando hidratos de carbono como agentes redutores:

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \frac{C}{R_p}\exp\Big[-\Big(\frac{E}{R}\Big(\frac{1}{T}-\frac{1}{T'}\Big)+\frac{b_1 X^{b_2}}{RT}\Big)\Big]\times \Big(c^0_{\mathrm{H_2SO_4}}-\frac{\nu_{\mathrm{MnO_2}}}{\nu_{\mathrm{H_2SO_4}}}c^0_{\mathrm{MnO_2}}X\Big)^a\Big(c^0_{\mathrm{C}}-\frac{\nu_{\mathrm{MnO_2}}}{\nu_{\mathrm{C}}}c^0_{\mathrm{MnO_2}}X\Big)^c(1-X)^{2/3} \quad (4.21)$$

onde X denota o grau de conversão definido em relação ao manganésio, t o tempo (min), C é uma constante (μm $(\mathrm{mol}^{-1}\ \mathrm{dm}^3)^{a+c}\ \mathrm{min}^{-1}$), R_p o tamanho médio das partículas de minério (μm), E a energia de activação (kJ/mol), R a constante dos gases (kJ/mol K), T a temperatura absoluta (K), T' é uma temperatura de referência (K), b_1 e b_2 são parâmetros que relacionam a conversão com a energia de activação (kJ/mol e adimensional, respectivamente), c^0 representa a concentração inicial de qualquer dos reagentes ($\mathrm{mol/dm}^3$), o índice C designa o hidrato de carbono, os expoentes a e c representam a ordem parcial de reacção em relação ao ácido e ao hidrato de carbono, respectivamente, e ν representa os coeficientes estequiométricos da reacção química global de extracção do metal. A temperatura de referência T' é dada por

$$T' = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\right)^{-1}$$

Repare-se que no modelo cinético acima é adicionado um termo de conversão ao termo exponencial de Arrhenius, que conduz ao aumento do valor da energia de activação com

Tabela 4.14 Lixiviação de minério manganífero: dados experimentais para a fracção 74 µm a 105 µm

				percentagem de manganésio extraído			
Teste	$[H_2SO_4]$/mol dm^{-3}	[Lactose]/g dm^{-3}	T/°C	5 min	15 min	30 min	60 min
1	0,5	10,0	30,0	2,0	4	8	14,6
2	0,5	30,0	30,0	2,2	8	13	22,3
3	1,5	10,0	30,0	2,4	7	12	23,9
4	1,5	30,0	30,0	3,4	12	20	49,6
5	0,5	10,0	70,0	17,4	35	51	65,0
6	0,5	30,0	70,0	27,7	52	64	73,6
7	1,5	10,0	70,0	38,3	63	75	86,2
8	1,5	30,0	70,0	46,8	74	81	86,2
9	1,0	36,8	50,0	17,9	39	53	68,8
10	1,0	3,2	50,0	6,1	16	31	48,8
11	1,8	20,0	50,0	14,1	37	60	73,2
12	0,2	20,0	50,0	6,6	17	30	43,9
13	1,0	20,0	83,6	45,1	78	82	88,0
14	1,0	20,0	16,4	2,4	7	13	21,5
15	1,0	20,0	50,0	10,8	30	48	65,2

a extensão da reacção. Este termo de "energia de activação variável", isto é $E + b_1 X^{b_2}$, descreve o decréscimo no rendimento de extracção de metal observado no decurso do tratamento de lixiviação.

Hidratos de carbono como a glucose e a lactose, entre outros, constituem alternativas mais seguras do ponto de vista ambiental do que os reagentes redutores convencionalmente utilizados no tratamento hidrometalúrgico de minérios de manganésio. Vegliò *et al.* (2001a) mostraram que o termo $1/R_p$ não é adequado para representar o efeito do tamanho das partículas no caso da lixiviação ácida do dióxido de manganésio com lactose como reagente redutor, embora o modelo cinético conduza a resultados satisfatórios quando aplicado separadamente a cada uma das três fracções granulométricas investigadas. Nessa situação estima-se $C' = C/R_p$ (mol^{-1} dm^3)$^{a+c}$ min^{-1}) e não C.

Os dados da tabela 4.14 dizem respeito a testes de lixiviação com lactose obtidos usando partículas de minério de dimensão compreendida entre 74 µm e 105 µm com 28,7 % em massa de dióxido de manganésio, na concentração de 30 g/dm^3. Neste caso $\nu_{MnO_2}/\nu_{H_2SO_4} = 1$ e $\nu_{MnO_2}/\nu_{\text{lactose}} = 1/24$.

Na utilização da rotina de integração `LSODA`, estabeleceu-se 10^{-6} para as tolerâncias relativa (RTOL) e absoluta (ATOL).

4.5.2 Aplicação aos dados experimentais

Os parâmetros de controlo do algoritmo MDE e as estimativas dos parâmetros do modelo (4.21) obtidas a partir dos dados experimentais constam das tabelas 4.15 na página seguinte e 4.16 na próxima página, respectivamente. O valor obtido para a estimativa do desvio padrão do erro de medição é de 0,04.

Observando a figura 4.4 na página seguinte, verifica-se que, grosso modo, as estimativas LS, L_p, e MM-LTS25 estão agrupadas. Esta característica é ainda partilhada pelas

Tabela 4.15 Lixiviação de minério manganífero: parâmetros do algoritmo de optimização MDE usados na regressão dos dados experimentais

Espaço de procura $[10^{-6}, 0, 1^{a}, 0{,}5, 0, 0] \leqslant [C', E, b_1, b_2, a, c] \leqslant [1, 100, 100, 10, 1^{a}, 1^{a}]$					
N_P	F_b	C_R	ξ	Tolerância	Ponto incluído na população inicial
31	0,5	0,8	1,5	10^{-8}	$[0{,}09, 59, 12, 2{,}9, 0{,}73, 0{,}51]^{b}$

[a] Estes valores foram tomados das restrições impostas por Vegliò *et al.* (2001b, p. 3897) aos parâmetros correspondentes no caso da lixiviação com glucose.
[b] Tabela 8 de Vegliò *et al.* (2001a, p. 173).

Tabela 4.16 Lixiviação de minério manganífero: estimativas para os parâmetros do modelo

	Parâmetros					
Estimadores	$C'/[(\text{mol}^{-1}\ \text{dm}^3)^{a+c}\ \text{min}^{-1}]$	E/kJ mol^{-1}	b_1/kJ mol^{-1}	b_2	a	c
LS	0,076	54	12	2,9	0,63	0,42
L_p	0,057	56	15	3,6	0,7	0,34
LMS	0,084	59	22	5	0,56	0,5
LTS25	0,062	63	18	3,9	0,47	0,37
LTS50	0,094	58	24	5,5	0,49	0,53
MM-LTS25	0,066	56	14	3,2	0,64	0,38
τ	0,059	64	15	3,4	0,41	0,35
LTD25	0,056	62	20	4,4	0,49	0,36
LTD50	0,063	63	16	3,7	0,48	0,39

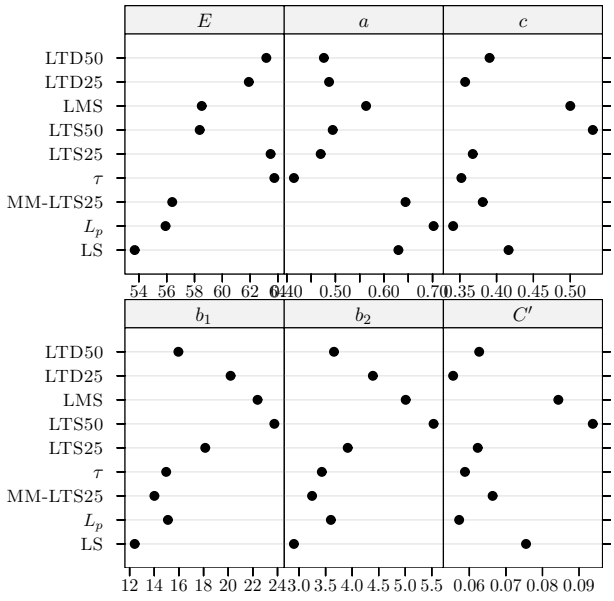


Figura 4.4 Lixiviação de minério manganífero: estimativas para os parâmetros do modelo.

Tabela 4.17 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: evolução temporal da densidade de células (x) para o nível de inoculação de 2,26 células/microcarrier. Dados extraídos da tabela 17.13 de Englezos e Kalogerakis (2001, p. 345)

Obs.	Tempo/h	$x/10^6$ células cm^{-3}
1	21,8	0,06
2	24,2	0,06
3	30,2	0,07
4	41,7	0,08
5	48,4	0,13
6	66,5	0,09
7	73,8	0,15
8	91,9	0,34
9	99,2	0,39
10	111,3	0,65
11	118,5	0,67
12	134,3	0,92
13	144,0	1,24
14	158,5	1,47
15	166,9	1,36
16	182,7	1,56
17	205,6	1,52
18	215,3	1,61
19	239,5	1,78

estimativas τ e LTS25, bem como pelas estimativas LMS e LTS50. Em geral, estas últimas apresentam previsivelmente a discrepância mais elevada em relação à estimativa LS.

4.6 Crescimento de células MRC-5 em microcarriers

4.6.1 Descrição genérica do problema

As células animais não são capazes de crescerem livres como os microrganismos. Por conseguinte, dependem da ancoragem a um suporte, isto é, requerem a ligação a uma superfície para o seu crescimento. Por isso, a cultura de tais células ocorre à superfície de suportes sólidos, sendo a proliferação de cada célula inibida pelo contacto com células adjacentes. No caso duma cultura em monocamada, quando toda a superfície estiver coberta o crescimento e divisão celular páram. Diz-se então que a cultura se encontra em confluência.

Os dados da tabela 4.17 dizem respeito a uma experiência relatada em Hawboldt *et al.* (1994), para estudar a inibição do crescimento de culturas de células MRC-5 em *microcarriers* — pequenas esferas de matriz sólida não porosa — com o objectivo de validar o modelo computacional baseado em autómatos celulares descrito nesse artigo. Aqui, a descrição do crescimento é dada pelo seguinte modelo proposto por Frame e Hu (1988)

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \mu_{\max}\Big[1 - \exp\Big(-C\frac{x_\infty - x}{x_\infty}\Big)\Big]x, \tag{4.22}$$

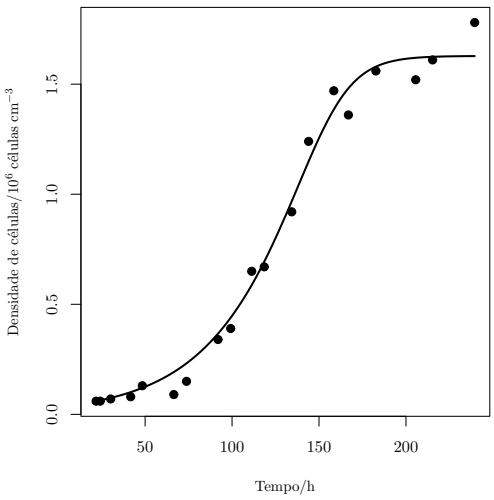


Figura 4.5 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: variação com o tempo da densidade de células, calculada com $\mu_{\max} = 0{,}0280$, $C = 2{,}86$, $x_\infty = 1/0{,}6142$ segundo o modelo (4.22). Os pontos representam os dados experimentais. Nível de inoculação de 2,26 células/microcarrier.

Tabela 4.18 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: parâmetros do algoritmo de optimização MDE usados na regressão dos dados experimentais

Espaço de procura $[0,0,1] \leqslant [\mu_{\max}, C, x_\infty] \leqslant [1,100,2{,}5]$					
N_{P}	F_{b}	C_{R}	ξ	Tolerância	Ponto incluído na população inicial
31	0,5	0,8	1,5	10^{-15}	$[0{,}0280, 2{,}86, 1/0{,}6142]$[a]

[a] Englezos e Kalogerakis (2001, p. 344).

onde x é a densidade média de células na cultura (10^6 células/cm^3), $\mu_{\max}$ é a taxa específica máxima de crescimento (h^{-1}), C é uma constante, e x_∞ denota a densidade máxima de células que pode ser atingida em confluência (10^6 células/cm^3). Para condição inicial toma-se a primeira observação.

Na utilização da rotina de integração `LSODA`, estabeleceu-se 10^{-6} para as tolerâncias relativa (RTOL) e absoluta (ATOL).

4.6.2 Aplicação aos dados experimentais

Na tabela 4.18 apresentam-se os parâmetros de controlo do algoritmo MDE. As estimativas obtidas para os parâmetros do modelo (4.22) a partir dos dados experimentais encontram-se na tabela 4.19 na próxima página. O valor obtido para a estimativa do desvio padrão do erro de medição é de $0{,}07815 \times 10^6$ células/cm^3. Na figura 4.6 na página ao lado apresentam-se gráficos para os resultados referentes à tabela 4.19 na próxima página.

De acordo com Englezos e Kalogerakis (2001, pp. 344–346) a precisão da estimativa LS de C é bastante pequena. No que diz respeito aos estimadores robustos, pode-se conjecturar ser válida uma conclusão semelhante, o que justifica a substancial variação das correspondentes estimativas apresentadas na tabela acima. É de destacar que as estimativas LMS e LTS50 têm valores bastante elevados em comparação com as restantes. Não parece haver diferenças significativas entre as estimativas robustas obtidas para

Tabela 4.19 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: estimativas para os parâmetros do modelo

	Parâmetros		
Estimadores	$\mu_{\max}/\mathrm{h}^{-1}$	C	$x_\infty/10^6$ células cm^{-3}
LS	0,028	2,86	1,63
L_p	0,0251	6,33	1,56
LMS	0,0247	24,1	1,54
LTS25	0,0247	8,27	1,56
LTS50	0,0248	19,5	1,52
MM-LTS25	0,0252	5,76	1,57
τ	0,025	7,87	1,53
LTD25	0,025	7,28	1,57
LTD50	0,0252	8,03	1,53

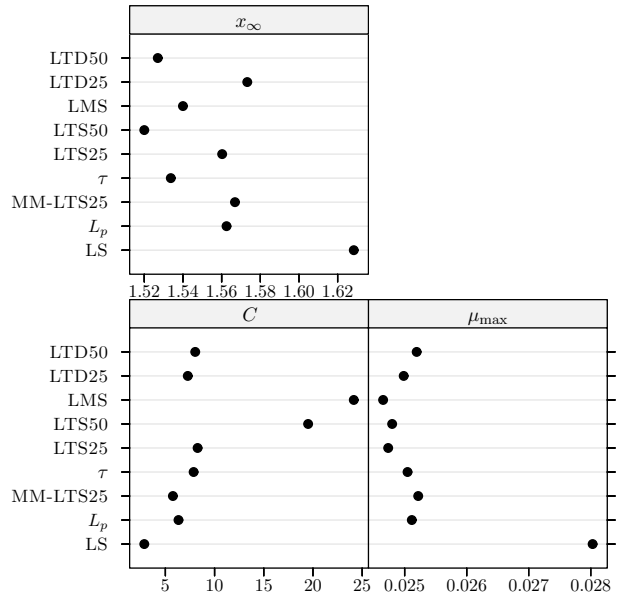


Figura 4.6 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: estimativas para os parâmetros do modelo.

Tabela 4.20 Tempos de cálculo dos diferentes problemas para o cenário sem *outliers*, obtidos com um processador Pentium 4 a 1,8 GHz e a versão R 1.4.0

Problema	Tempo de cálculo/h
Isomerização do *n*-pentano	10,1
Oxidação do propeno	15,3
Equilíbrio líquido-vapor	78,4
Equilíbrio sólido-líquido	5,6
Lixiviação de minério manganífero	102,8
Crescimento de células MRC-5	16,2

os parâmetros $\mu_{\max}$ e x_∞. Observando novamente a tabela verifica-se que o valor da estimativa LS de x_∞ é maior que os das estimativas robustas, embora a diferença seja pequena. A razão disto resulta da influência que a última observação (veja-se a figura 4.5 na página 62) exerce sobre o estimador LS — neste caso a fase estacionária da curva foi atraída na direcção dessa observação e em consonância o valor de x_∞ foi "empurrado" para cima — mas a que os estimadores robustos são insensíveis.

4.7 Resultados e discussão das experiências com dados simulados sem outliers

Para uma percepção do esforço computacional envolvido nas experiências de Monte Carlo apresentam-se na tabela 4.20, para os diversos problemas estudados, os tempos de cálculo requeridos no cenário actual.

Nota Como já se referiu, a qualidade estatística das estimativas LS de β_1 no caso da isomerização catalítica do *n*-pentano é pobre, e parece razoável assumir que esta característica se verifica igualmente com outros estimadores. É, pois, natural antever que a análise deste parâmetro se revele problemática e inconclusiva, que é o que acontece. De facto, a observação das figuras relativas a este caso nas subsecções adiante revela um padrão dos índices de desempenho estranho em relação a β_1, claramente diferente do dos outros parâmetros. Consequentemente, optou-se neste trabalho por omitir a análise deste parâmetro.

Nas figuras 4.7 a 4.12 nas páginas 65–69 pode observar-se o que se passa com a medida de eficiência robusta, e nas figuras 4.13 a 4.18 nas páginas 70–74 encontra-se o índice RB normalizado.

Critério de eficiência Considerando primeiro as distribuições Gaussiana, CN(0,10, 2), e Gaussiana enviesada, observe-se que o padrão de resultados de cada uma delas é semelhante. Mais concretamente e como seria de esperar, as estimativas LS apresentam em muitos casos o melhor desempenho sob a distribuição Gaussiana.[2] É interessante notar, no entanto, que o mesmo acontece com CN(0,10, 2) e a distribuição Gaussiana

[2]Note-se que embora para erro Gaussiano o estimador LS ser estimador assimptoticamente mais eficiente (Seber e Wild, 1989, p. 33), em amostras finitas podem, em princípio, existir estimadores mais eficientes do que o estimador LS.

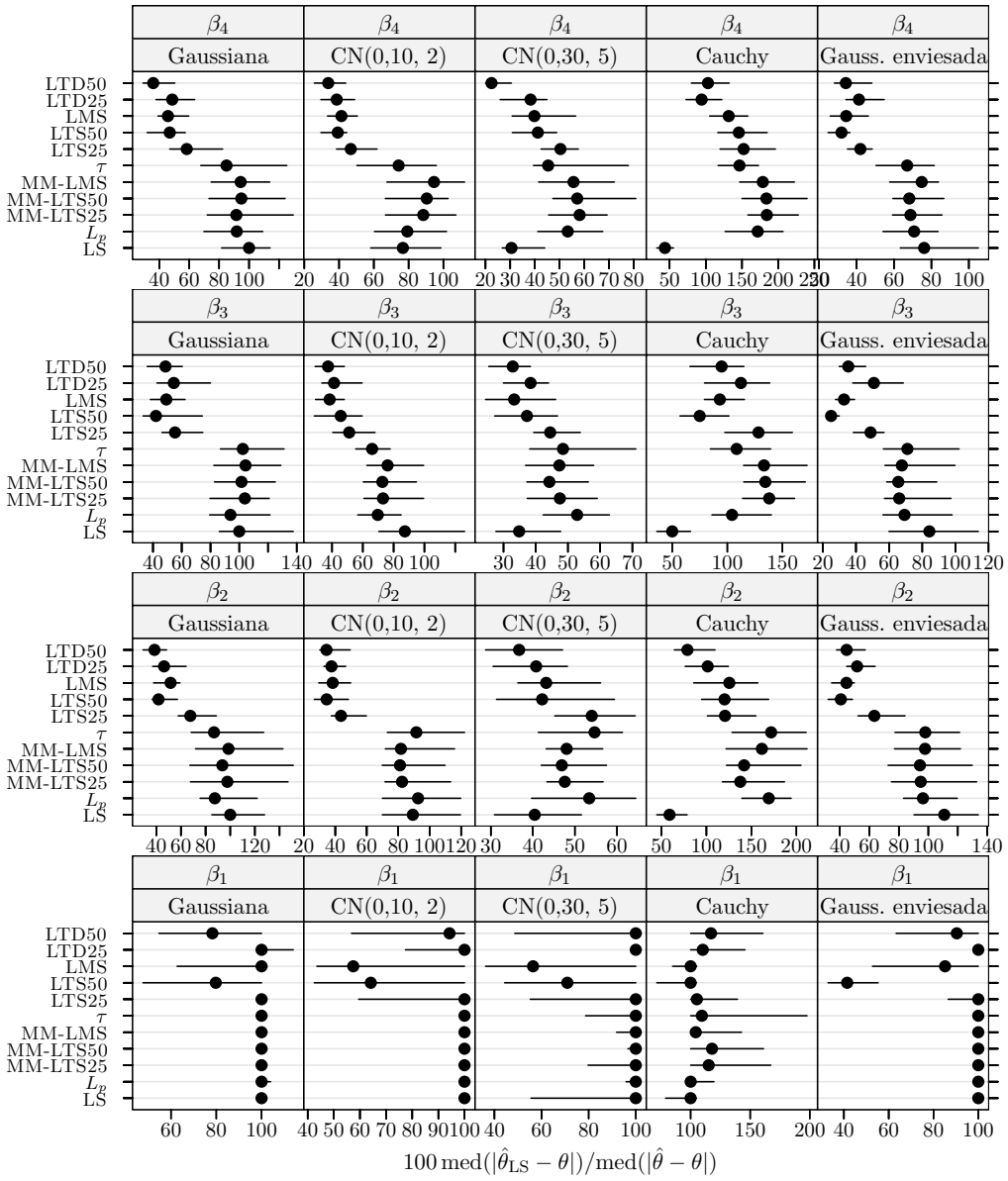


Figura 4.7 Isomerização do *n*-pentano: medida da eficiência dos estimadores para dados simulados sem *outliers*. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

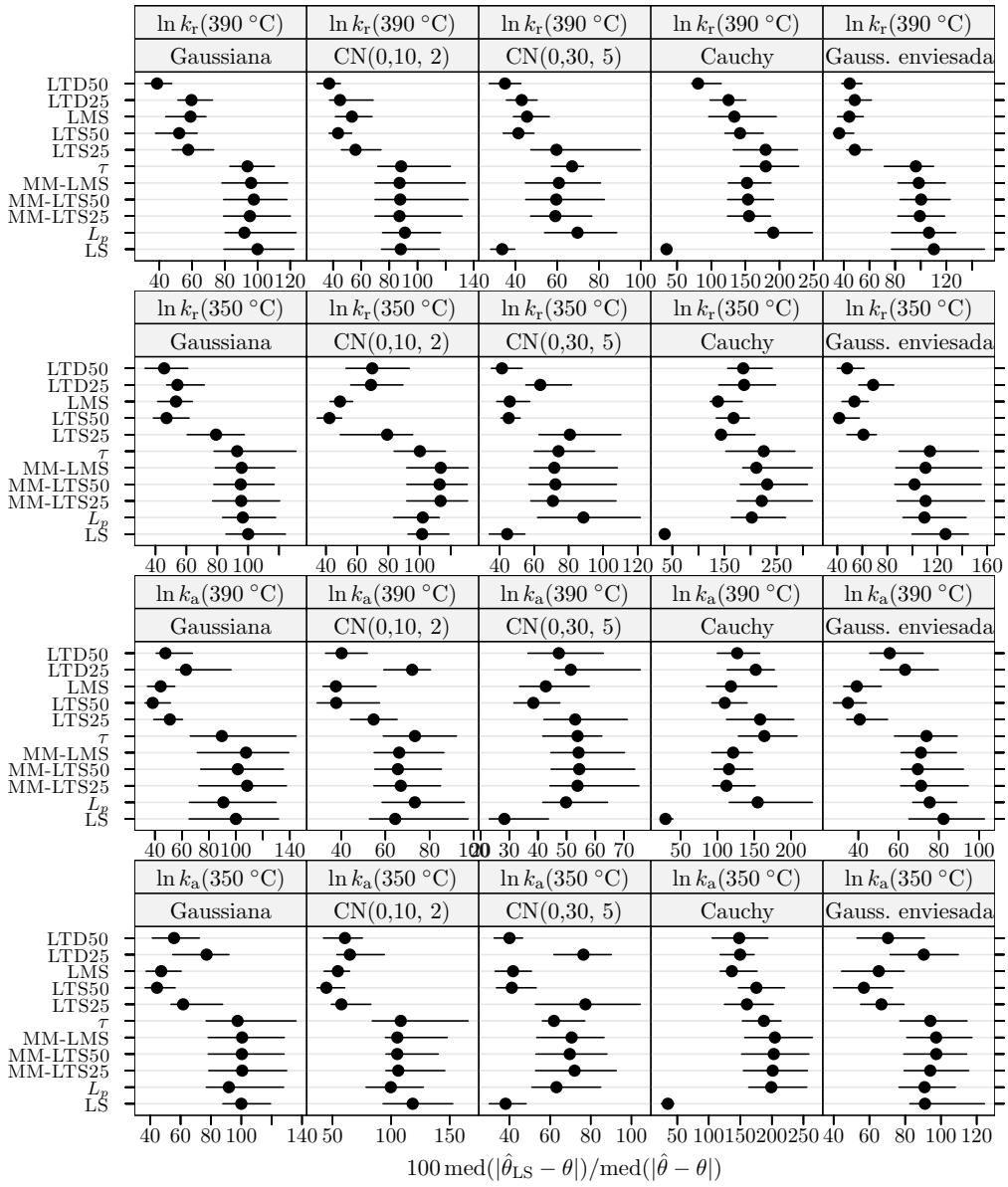


Figura 4.8 Oxidação do propeno: medida da eficiência dos estimadores para dados simulados sem *outliers*. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

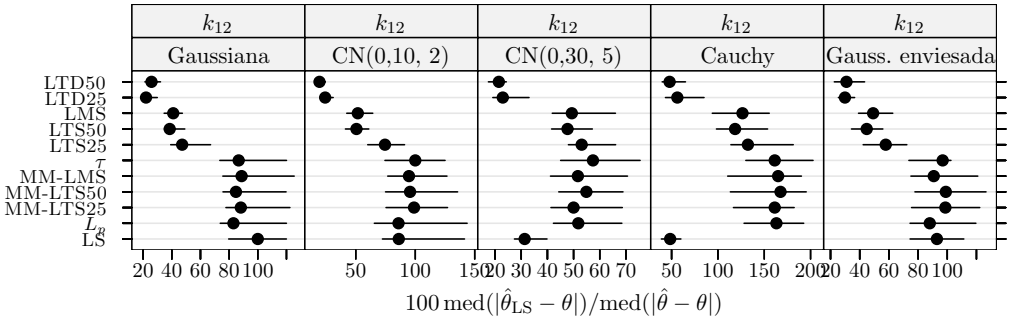


Figura 4.9 Equilíbrio líquido-vapor do sistema $CH_3OCH_3(1)$ + $CF_3CH_2CF_3(2)$: medida da eficiência dos estimadores para dados simulados sem *outliers*. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

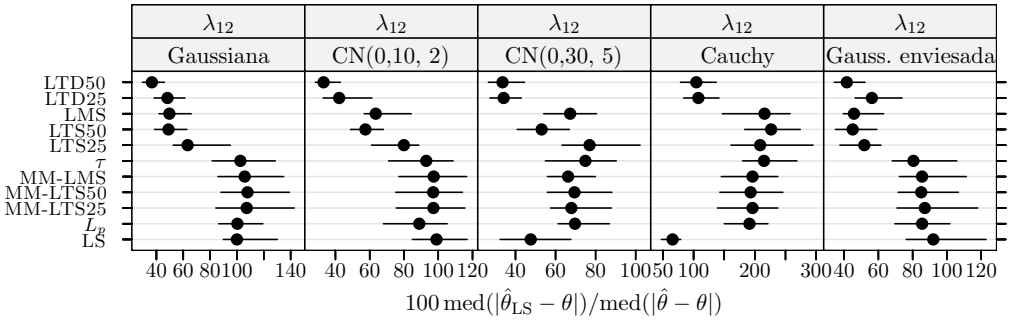


Figura 4.10 Equilíbrio sólido-líquido do sistema 1,4-butanodiol(1) + 4-metoxifenol(2): medida da eficiência dos estimadores para dados simulados sem *outliers*. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

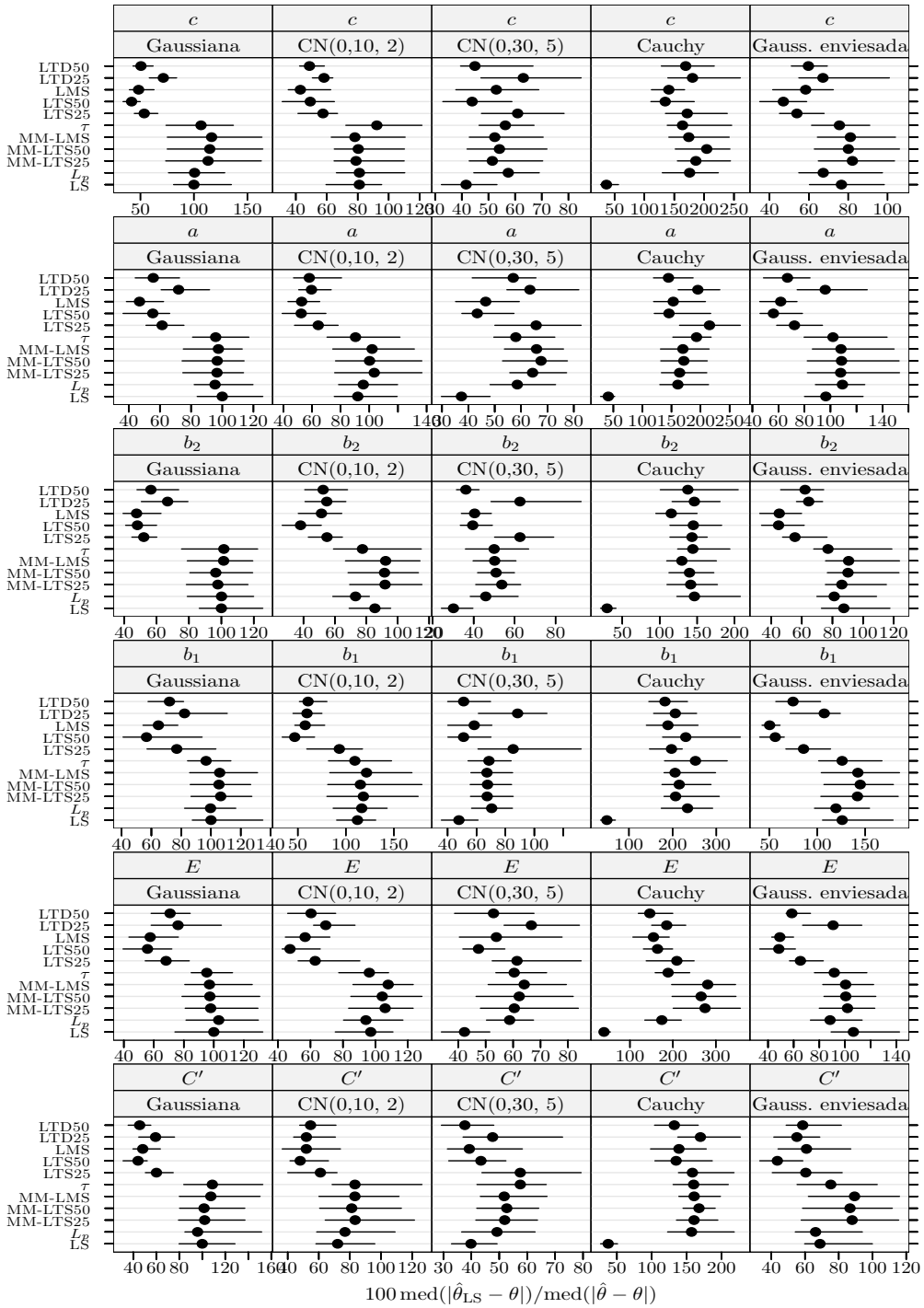


Figura 4.11 Lixiviação de minério manganífero: medida da eficiência dos estimadores para dados simulados sem *outliers*. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

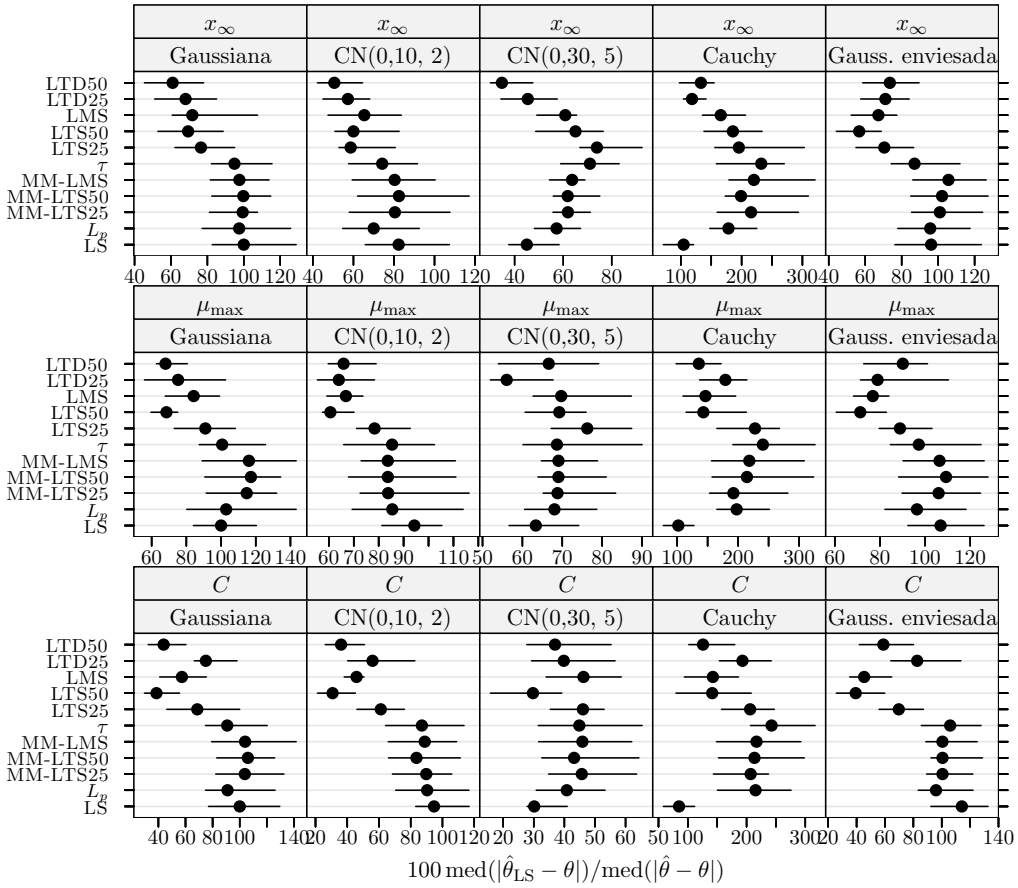


Figura 4.12 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: medida da eficiência dos estimadores para dados simulados sem *outliers*. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

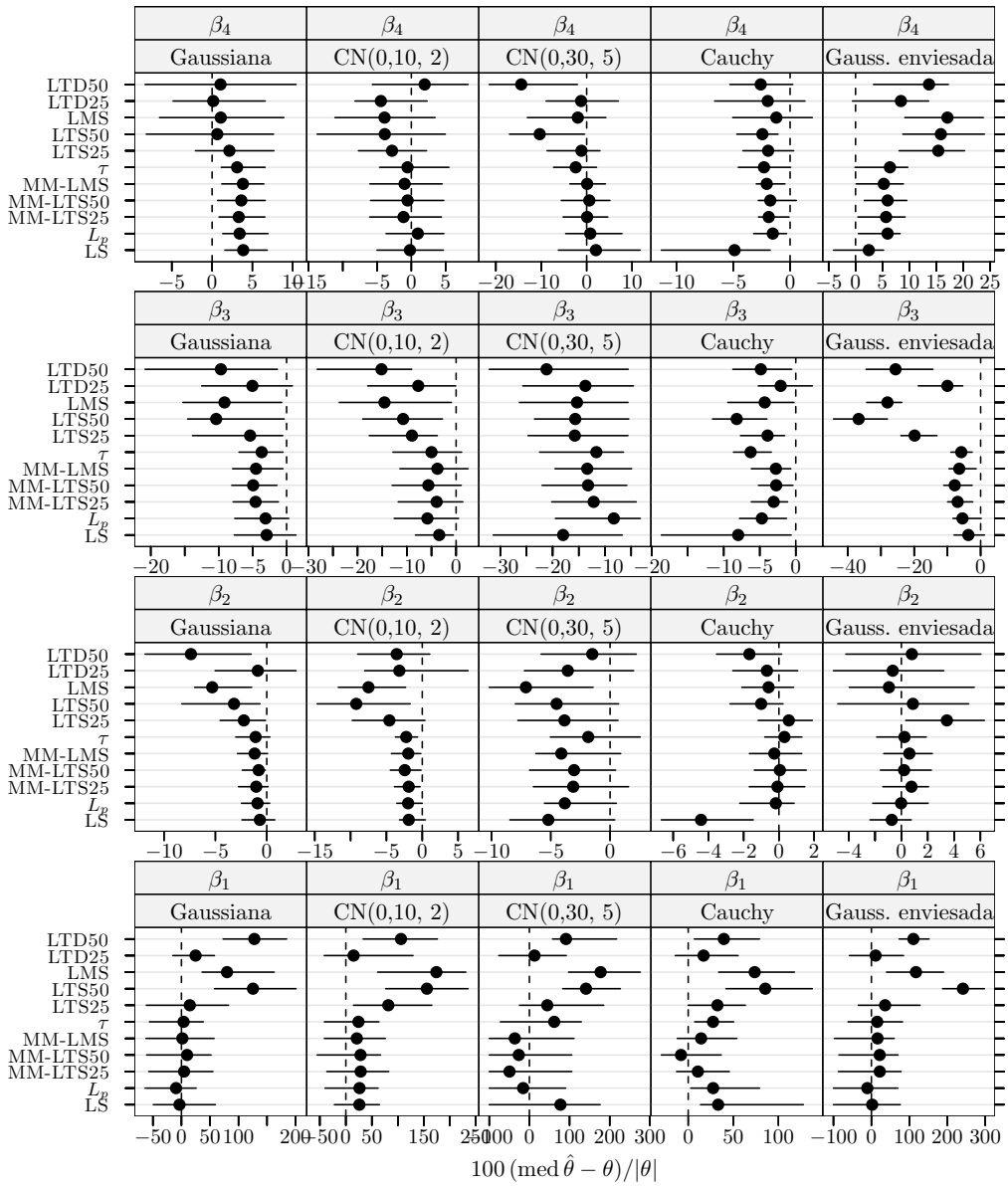


Figura 4.13 Isomerização do n-pentano: índice de enviesamento robustificado dos estimadores para dados simulados sem *outliers*. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

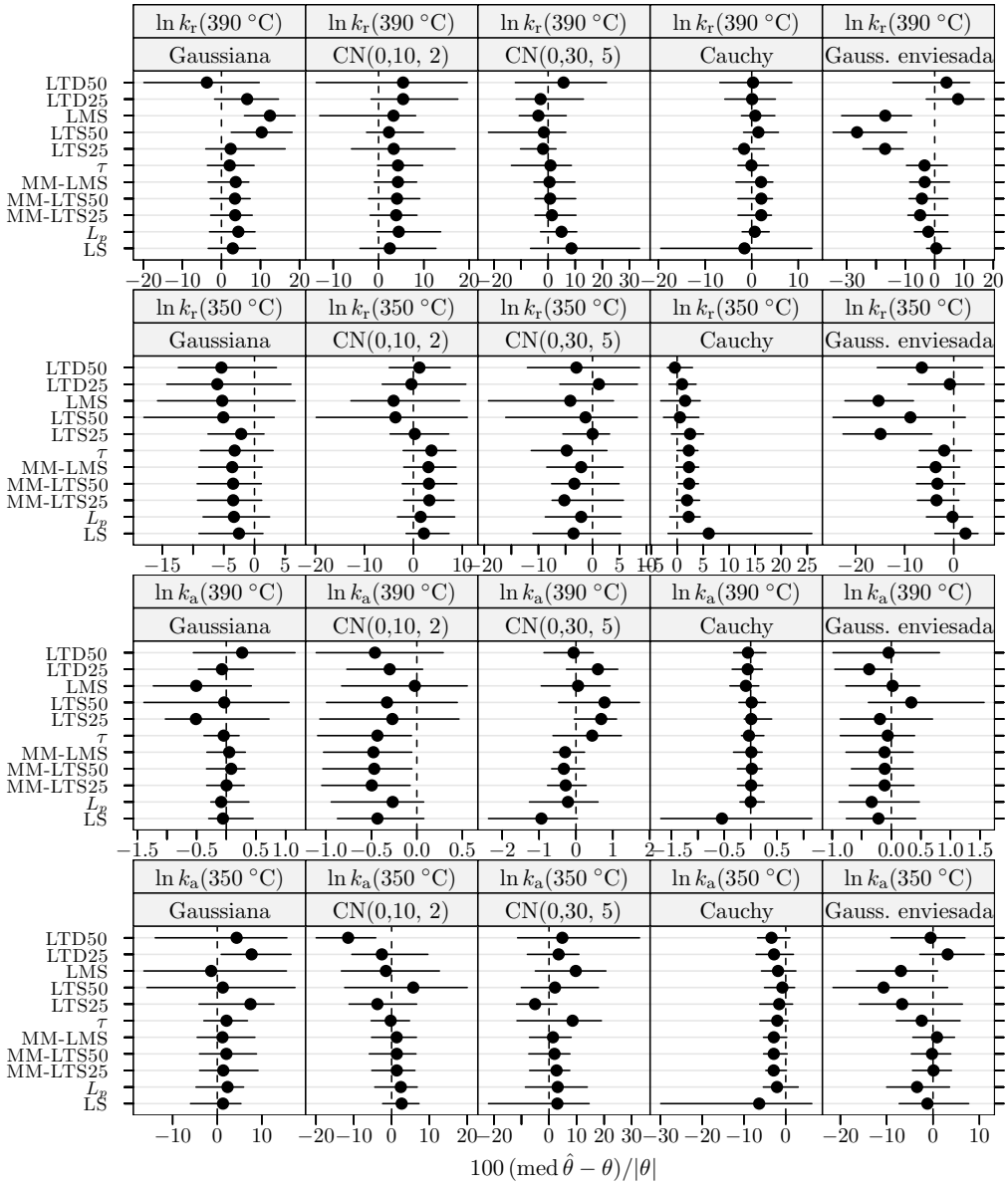


Figura 4.14 Oxidação do propeno: índice de enviesamento robustificado dos estimadores para dados simulados sem *outliers*. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

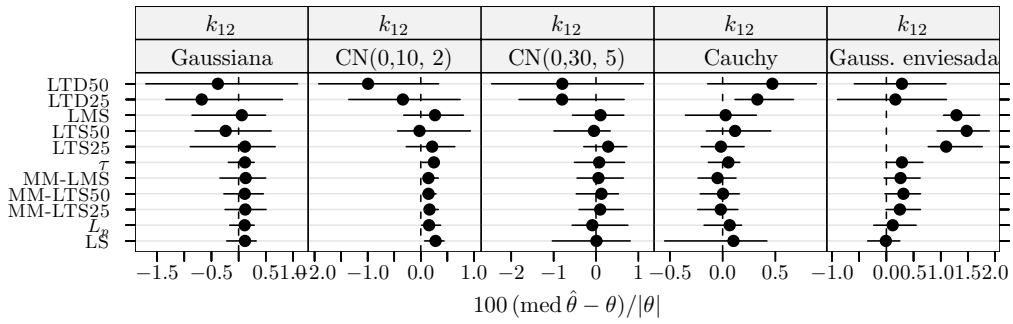


Figura 4.15 Equilíbrio líquido-vapor do sistema $CH_3OCH_3(1)$ + $CF_3CH_2CF_3(2)$: índice de enviesamento robustificado dos estimadores para dados simulados sem *outliers*. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

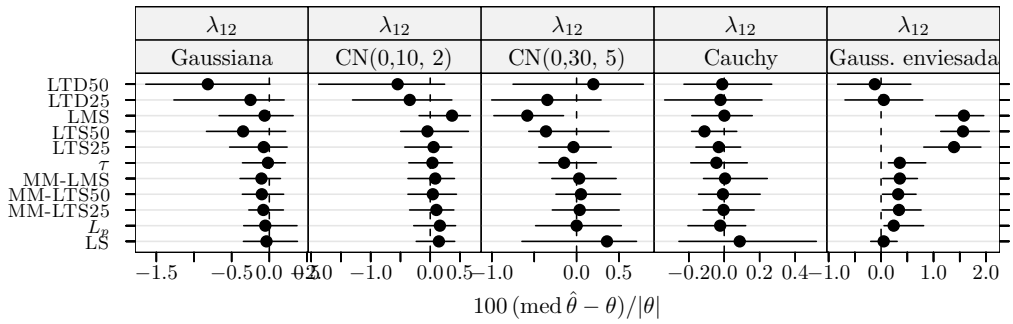


Figura 4.16 Equilíbrio sólido-líquido do sistema 1,4-butanodiol(1) + 4-metoxifenol(2): índice de enviesamento robustificado dos estimadores para dados simulados sem *outliers*. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

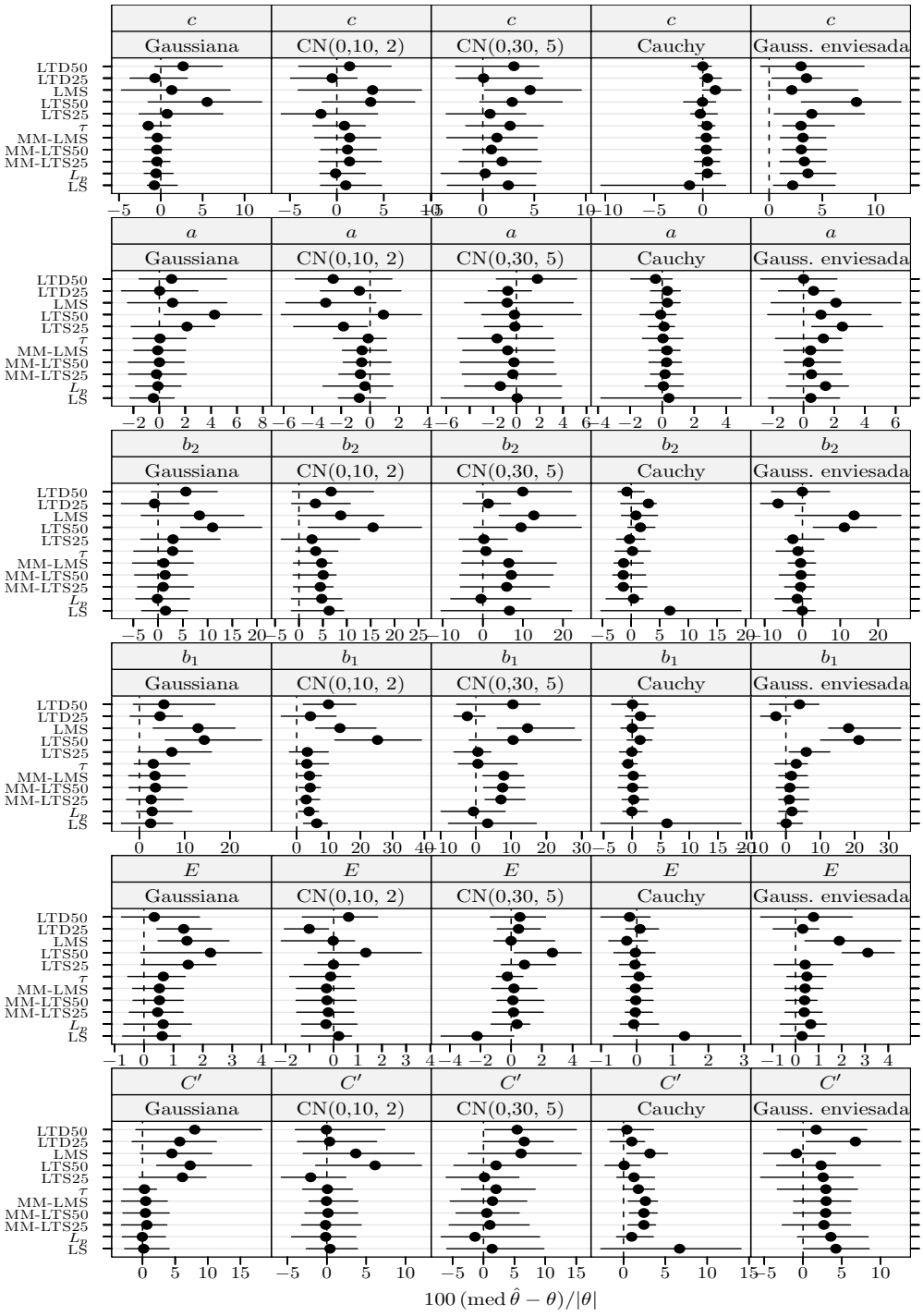


Figura 4.17 Lixiviação de minério manganífero: índice de enviesamento robustificado dos estimadores para dados simulados sem *outliers*. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

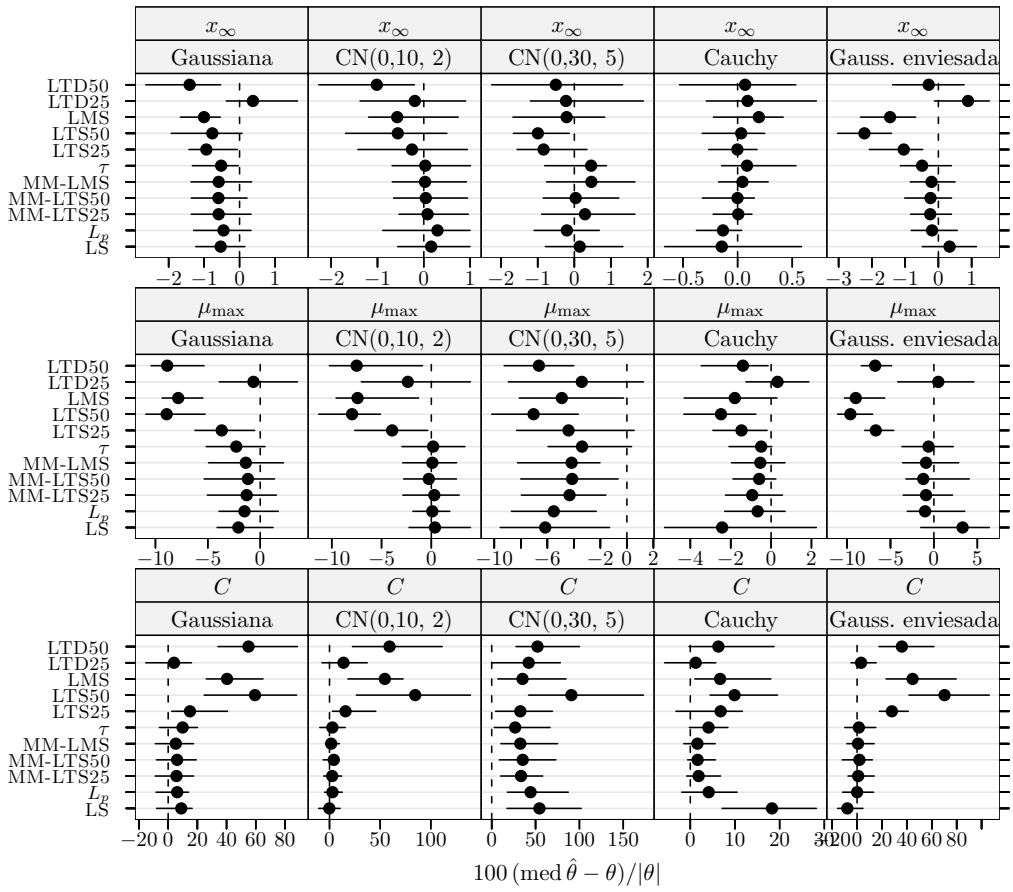


Figura 4.18 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: índice de enviesamento robustificado dos estimadores para dados simulados sem *outliers*. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

enviesada, ou dito de outra forma, não parece haver degradação do desempenho do estimador LS face a desvios moderados da assunção de erro de medição Gaussiano.

Adicionalmente, podem distinguir-se dois grupos de estimadores robustos, I e II, consoante o nível de desempenho: um composto pelos estimadores MM, τ, e L_p (I) e outro que inclui os estimadores LMS, LTS, e LTD (II). Como se pode observar há, de um modo consistente, uma clara superioridade do primeiro em relação ao segundo. A este propósito facilmente se reconhece que geralmente não existem diferenças significativas dentro dos grupos — em particular, as três variantes do estimador MM são essencialmente indistinguíveis —, bem como entre o grupo I e o estimador LS.[3] Em consonância, é importante sublinhar que os resultados não indiciam perdas de eficiência apreciáveis dos estimadores do grupo I sob a distribuição Gaussiana. Relativamente aos estimadores LTS e LTD, constata-se, sem surpresa, a superioridade da variante com 25% de aparamento sobre a variante com 50%, embora na maioria dos casos não haja diferença significativa entre estas versões.

Considerando agora as distribuições de Cauchy e CN(0,30, 5), constata-se um esbatimento na diferença de desempenho dos grupos I e II; é de destacar que nos casos de estudo de equilíbrio de fases o estimador LTD apresenta um desempenho pior do que o de todos os outros estimadores robustos. Além disso, com estas distribuições pode constatar-se uma forte deterioração no desempenho do estimador LS em relação aos melhores estimadores robustos; por outro lado, os estimadores LTS25, MM, τ, e L_p apresentam um desempenho globalmente muito satisfatório.

Por último, observe-se que o estimador LTD constitui uma desilusão, nomeadamente o pobre desempenho exibido quando sujeito a erro de medição assimétrico (caso da distribuição Gaussiana enviesada). Recorde-se que ao contrário dos restantes estimadores, o estimador LTD foi desenvolvido com o objectivo de acomodar distribuições de erro assimétricas.

Critério de enviesamento Por análise das figuras 4.13 a 4.18, observa-se que os valores de enviesamento são todos baixos ou moderados, excepto em relação ao parâmetro C do modelo de crescimento de células animais em *microcarriers*. Na maioria dos casos os diferentes estimadores são virtualmente indistinguíveis.

4.8 Resultados e discussão das experiências com dados simulados com outliers

Começa por considerar-se o cenário que pretende traduzir uma situação de contaminação moderada por *outliers*. Depois estudam-se dois cenários que pretendem representar situações de contaminação severa e posteriormente aborda-se o cenário que contempla circunstâncias extremas.

Observe-se que atendendo ao elevado grau de exigência usado na validação de dados experimentais em termodinâmica, não parece razoável considerar a contaminação com

[3] Quando os intervalos de confiança obtidos incluem valores dos índices de desempenho correspondentes a vários outros estimadores isto significa que não há evidência suficiente de diferença de desempenho entre esses métodos.

outliers para os casos de regressão de dados de equilíbrio de fases. Deste modo, aqueles são omitidos da presente secção.

4.8.1 Caso de contaminação moderada (10% de outliers, $\delta_{\mathsf{R}} = 5$)

Critério de eficiência A observação das figuras 4.19 a 4.22 nas páginas 77–80 revela um padrão de resultados que cabe no quadro que acabou de descrever-se para o cenário sem *outliers*. A comparação com o cenário anterior põe em evidência três aspectos principais:

- O primeiro aspecto é a significativa perda de eficiência dos estimadores robustos MM, τ, e L_p e do estimador LS. Em contraste, o mesmo não se passa com os restantes estimadores, os quais basicamente mantêm constante a eficiência. Consequentemente, aqui a clivagem existente entre o desempenho dos subconjuntos I e II de estimadores robustos atenua-se. Em particular, o desempenho do estimador LTS25 é em muitas situações comparável ao dos estimadores robustos mais eficientes. Uma leitura possível deste resultado aponta para o compromisso robustez/eficiência referido na secção 1.4 na página 8. Quer dizer, neste caso, a elevada eficiência atingida por estes estimadores sob a distribuição Gaussiana reflecte-se numa redução de robustez que conduz à perda de eficiência observada.

- O segundo é que as estimativas LS obtidas para alguns parâmetros dos vários modelos estudados (β_2 no modelo de isomerização do n-pentano, $\ln k_{\mathrm{r}}(390\ ^{\circ}\mathrm{C})$ e $\ln k_{\mathrm{r}}(350\ ^{\circ}\mathrm{C})$ no modelo de oxidação do propeno, bem como E e C' no modelo de lixiviação de minério manganífero) são aqui fortemente afectadas sob as distribuições Gaussiana, CN(0,10, 2), e Gaussiana enviesada, onde o estimador LS apresenta o valor mais baixo de eficiência de todos os estimadores.

- Uma terceiro aspecto é que sob a distribuição CN(0,30, 5) existe alguma evidência de uma ligeira deterioração do desempenho do estimador MM (particularmente notória no caso da variante MM-LTS25) em relação ao do estimador empregue no passo inicial. Novamente, isto pode ser interpretado como consequência da elevada eficiência (95%) sob distribuição Gaussiana deste estimador.

Critério de enviesamento Pela análise das figuras 4.23 a 4.26 nas páginas 81–84, observa-se que em geral o enviesamento dos diversos estimadores se acentua, salientando-se nitidamente o caso do estimador LS. É de referir que este último apresenta valores elevados de enviesamento para C' no modelo de lixiviação de minério manganífero, bem como para $\ln k_{\mathrm{r}}(390\ ^{\circ}\mathrm{C})$ e $\ln k_{\mathrm{r}}(350\ ^{\circ}\mathrm{C})$ no modelo de oxidação do propeno. Isto sugere que seja essencialmente o aumento do enviesamento que arrasta o forte decrescimento da eficiência das estimativas LS. Em oposição, a magnitude do enviesamento dos vários estimadores robustos permanece baixa ou moderada à excepção do parâmetro C do modelo de crescimento de células animais em *microcarriers*.

4.8.2 Caso de contaminação severa (20% de outliers, $\delta_{\mathsf{R}} = 5$)

Critério de eficiência Nas figuras 4.27 a 4.30 nas páginas 85–88 apresentam-se gráficos do índice de eficiência robustificado para os estimadores em competição. Neste cenário não é manifestamente fácil definir um quadro de referência simples.

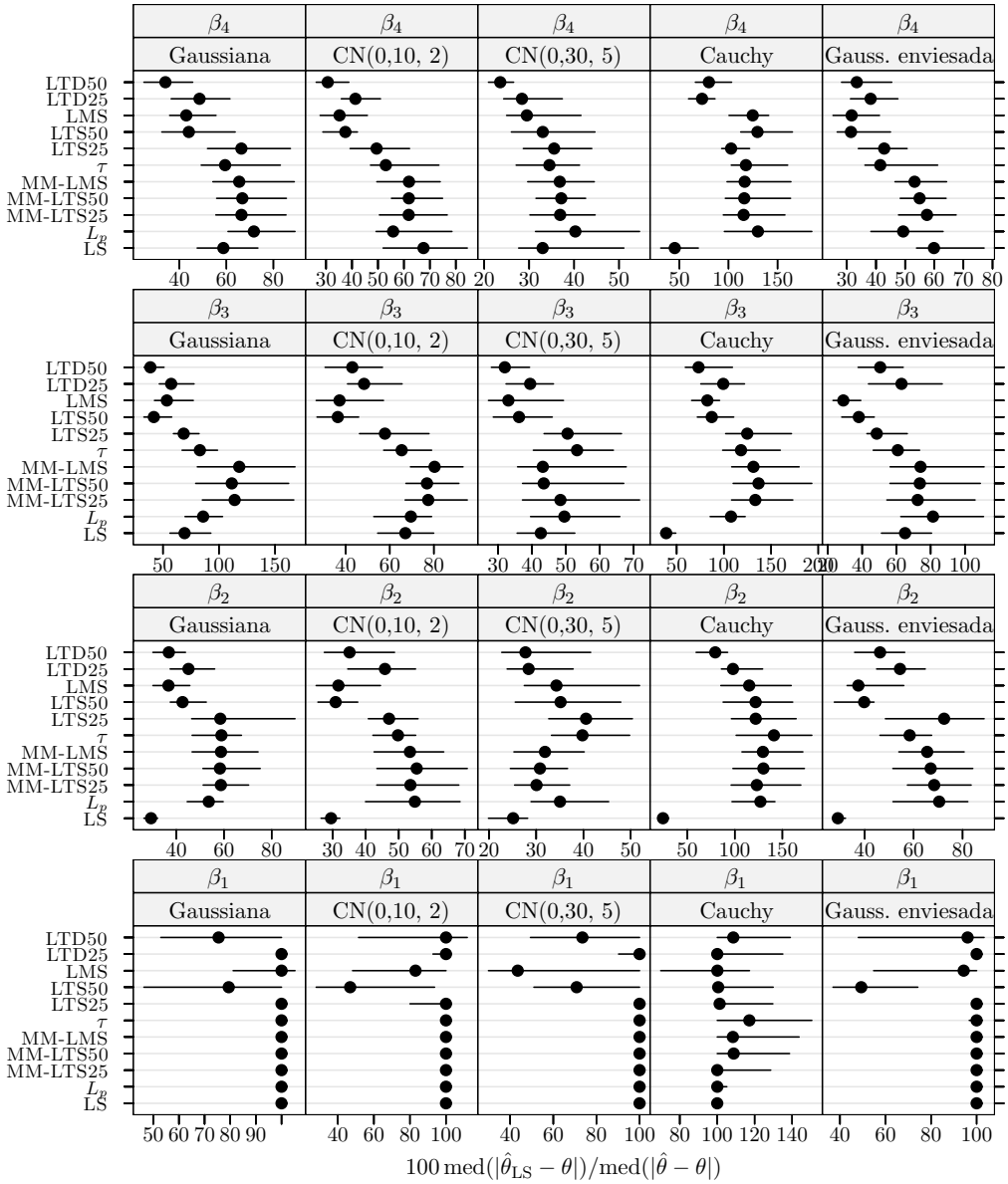


Figura 4.19 Isomerização do *n*-pentano: medida da eficiência dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 13 e 15.

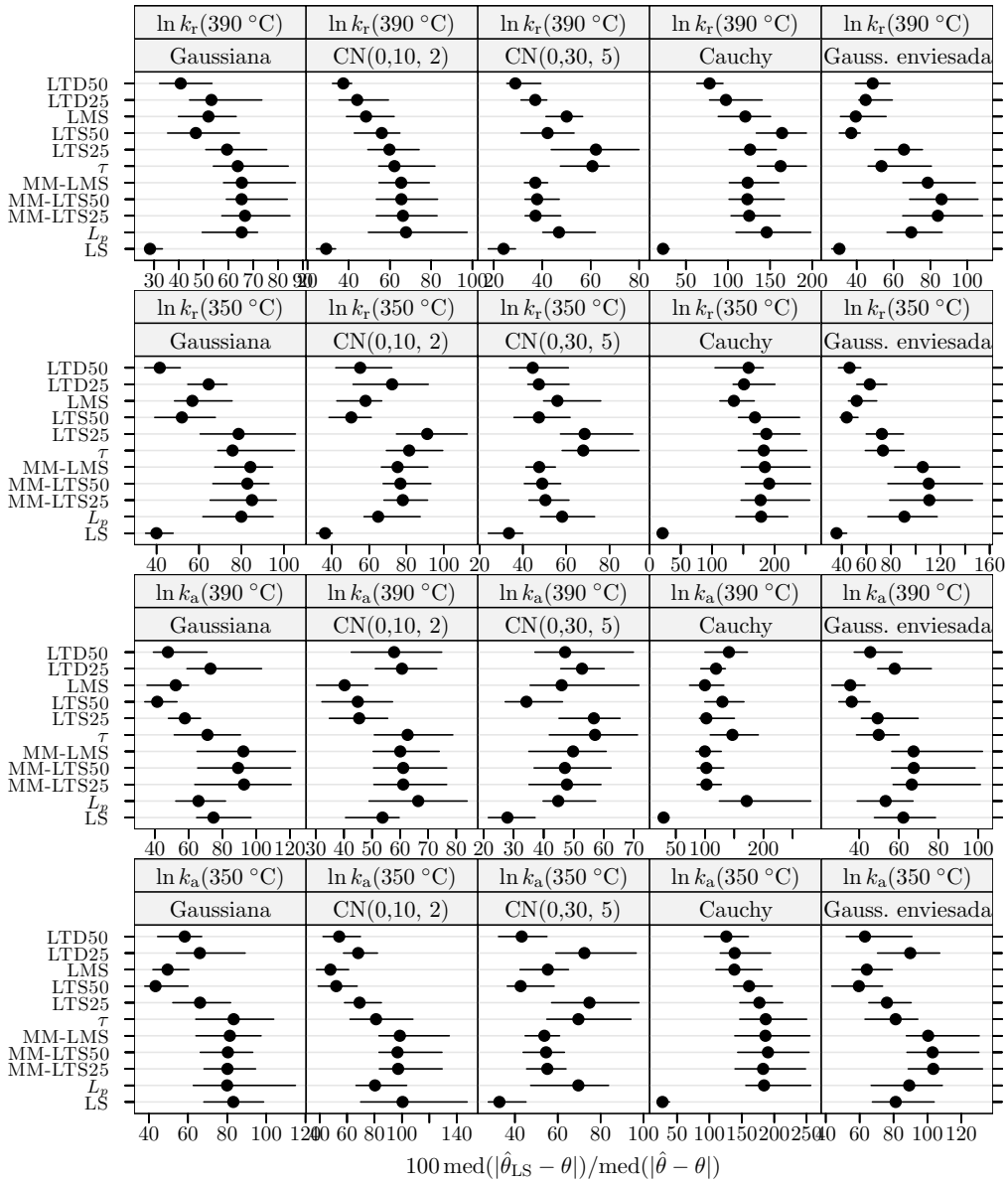


Figura 4.20 Oxidação do propeno: medida da eficiência dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 8, 24, 32, 33, 39, 47, e 49.

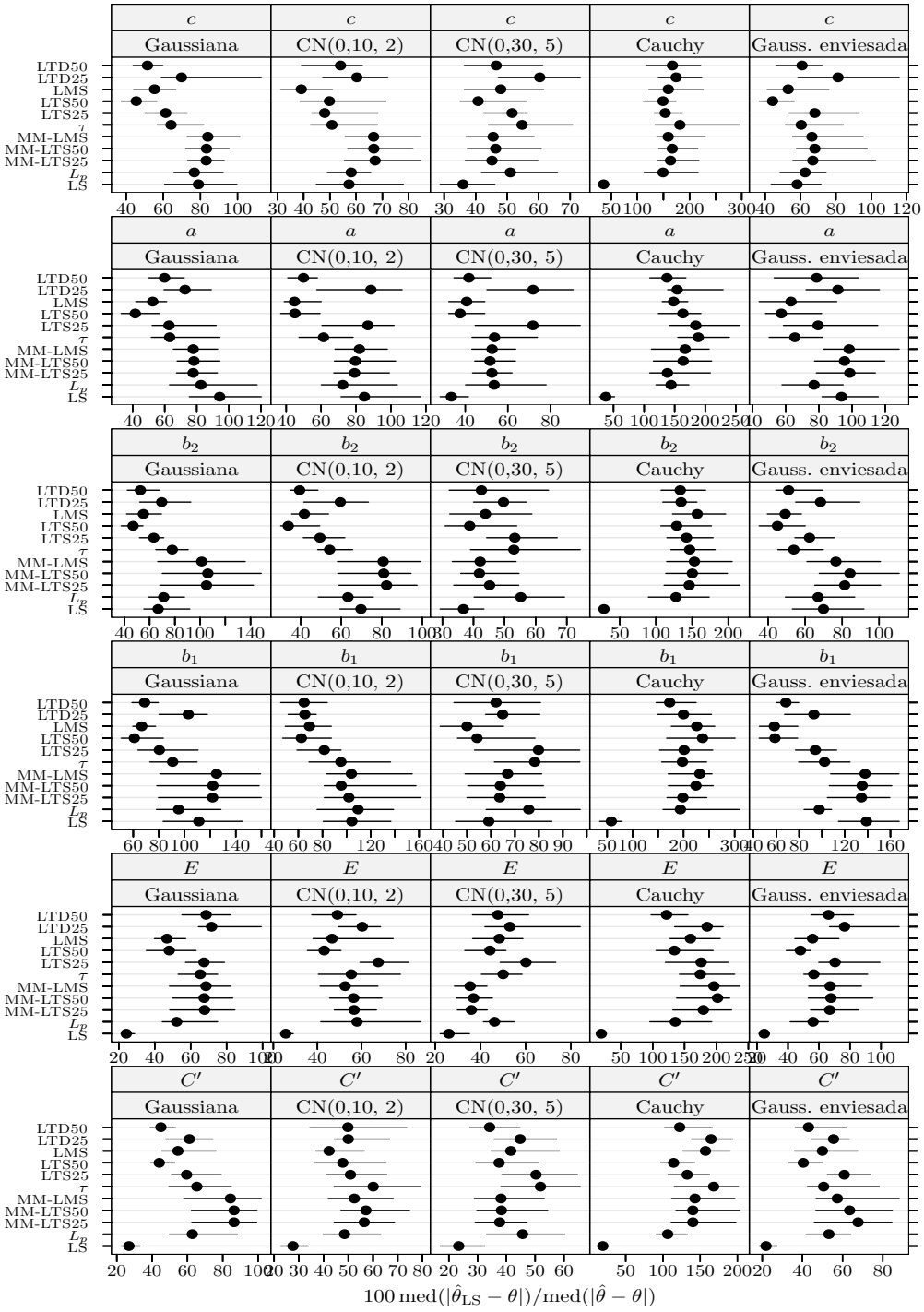


Figura 4.21 Lixiviação de minério manganífero: medida da eficiência dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 3, 17, 18, 29, 33, e 60.

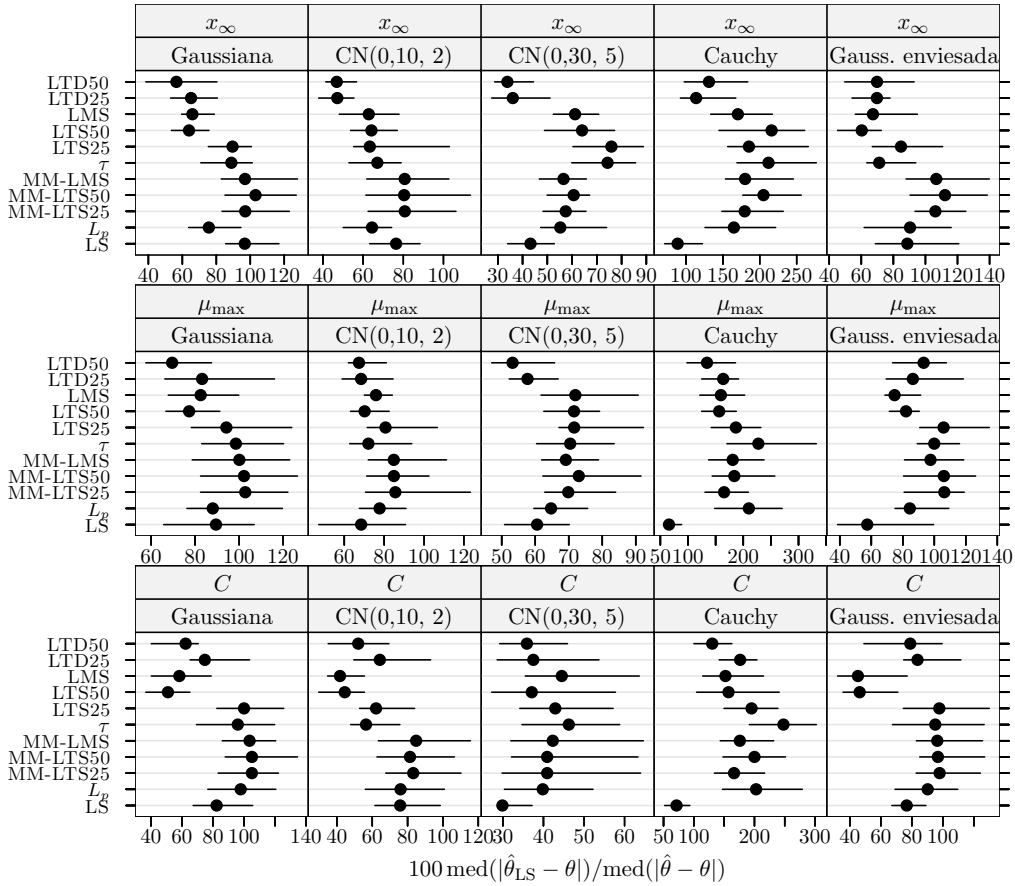


Figura 4.22 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: medida da eficiência dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1 e 5.

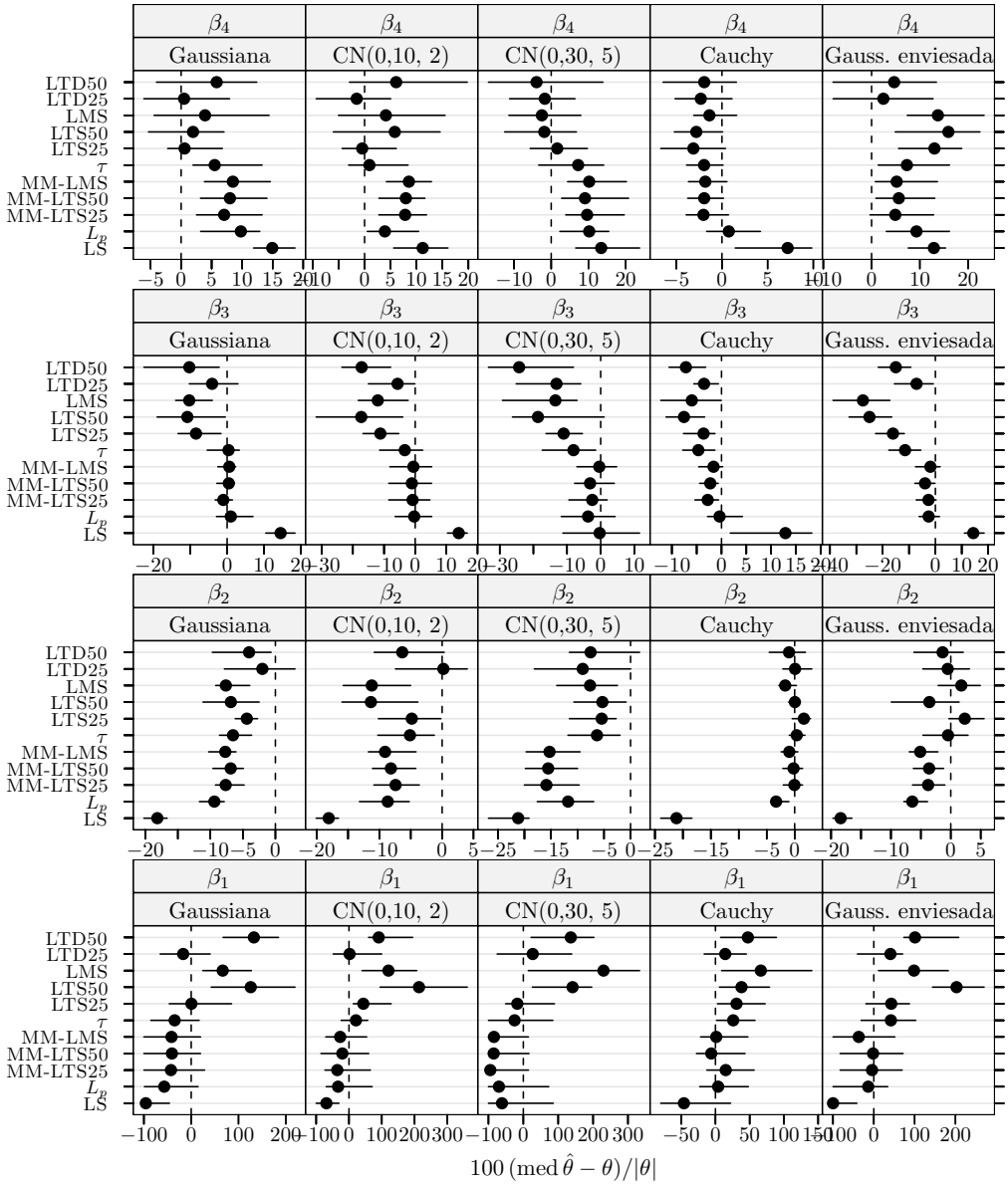


Figura 4.23 Isomerização do *n*-pentano: índice de enviesamento robustificado dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 13 e 15.

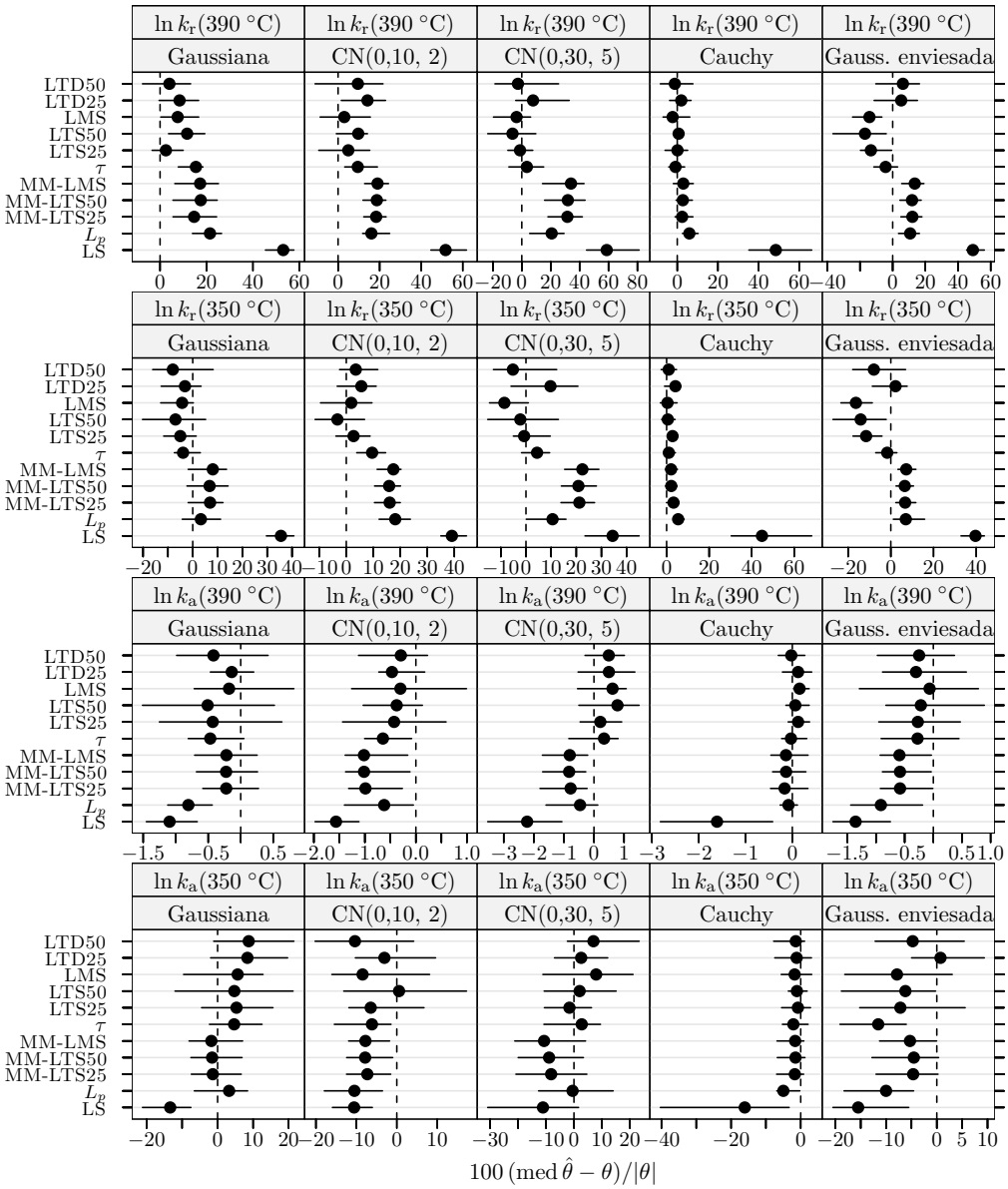


Figura 4.24 Oxidação do propeno: índice de enviesamento robustificado dos estimadores para dados simulados com 10% de *outliers* e $\delta_R = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 8, 24, 32, 33, 39, 47, e 49.

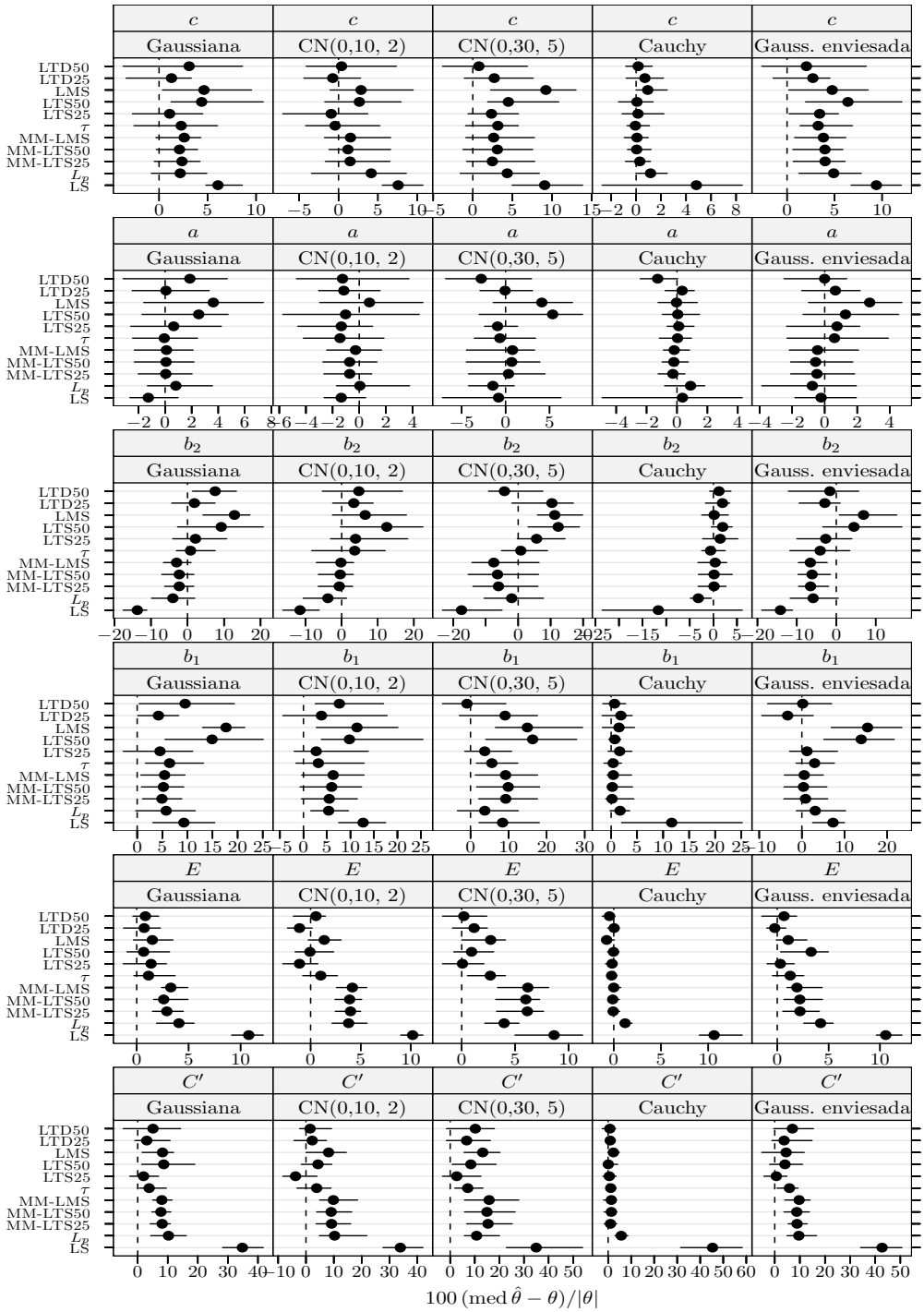


Figura 4.25 Lixiviação de minério manganífero: índice de enviesamento robustificado dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 3, 17, 18, 29, 33, e 60.

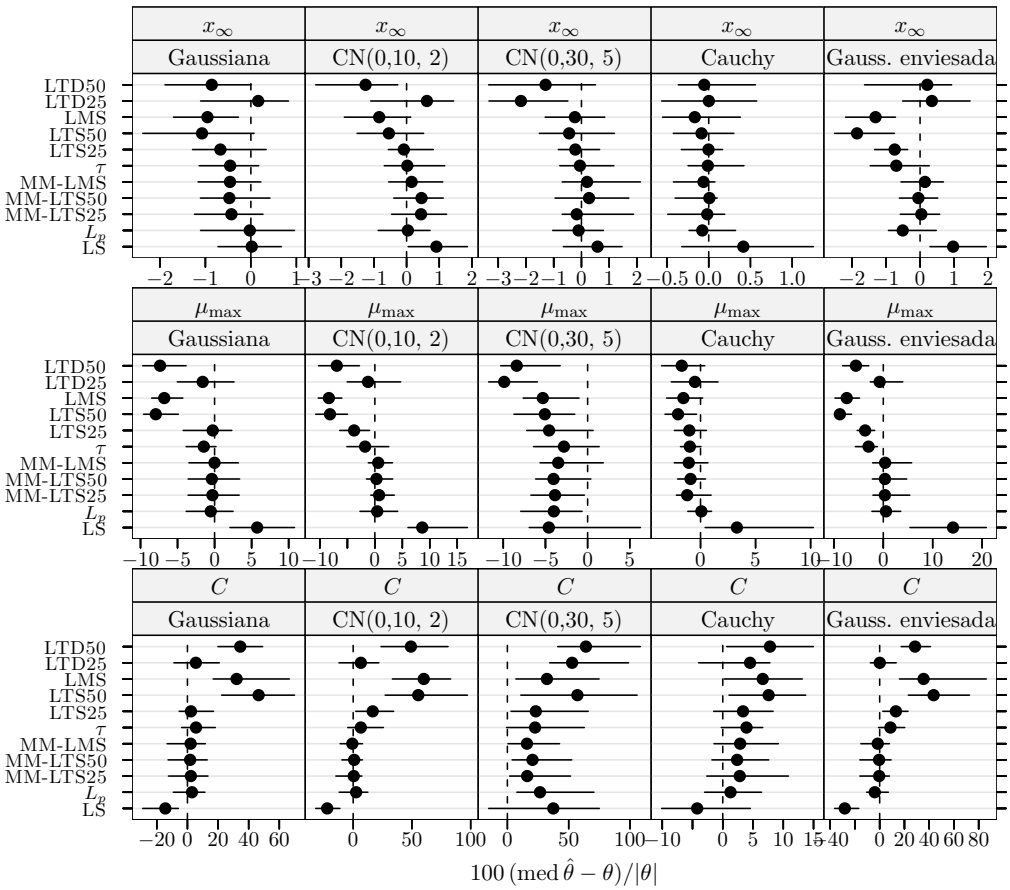


Figura 4.26 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: índice de enviesamento robustificado dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1 e 5.

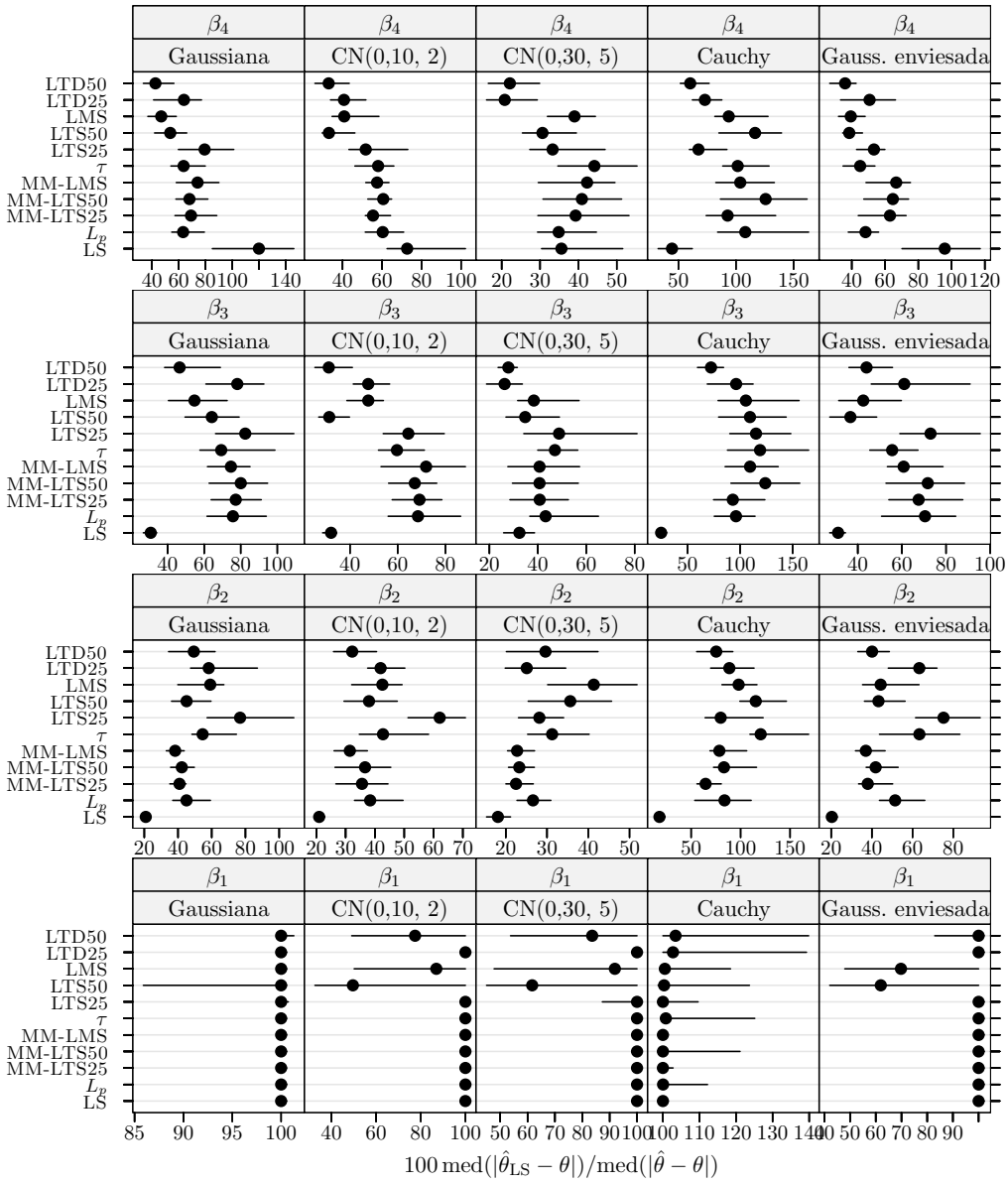


Figura 4.27 Isomerização do *n*-pentano: medida da eficiência dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 11, 13, 14, 15, e 21.

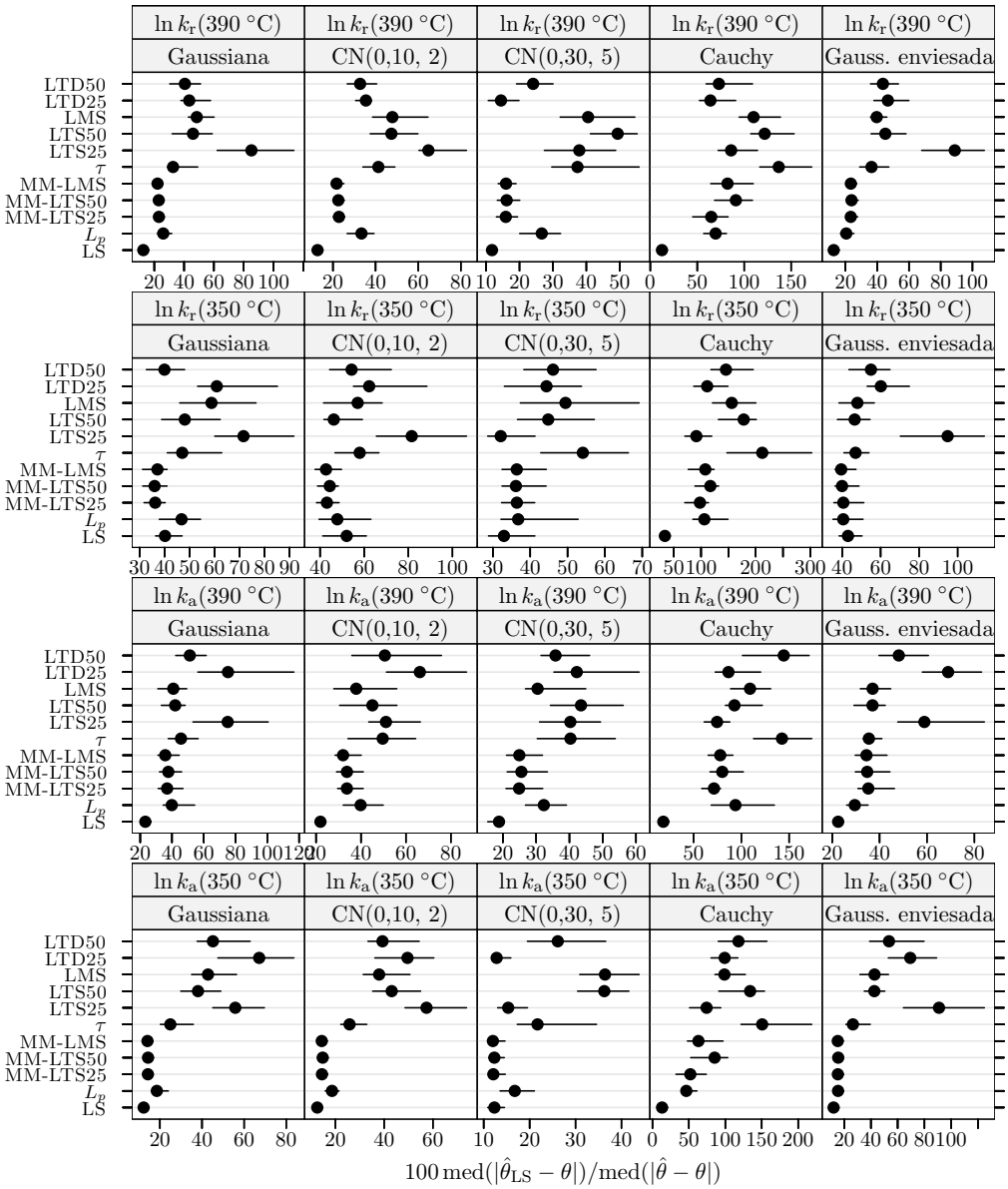


Figura 4.28 Oxidação do propeno: medida da eficiência dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 5, 8, 17, 19, 24, 32, 33, 39, 45, 47, 49, 56, e 66.

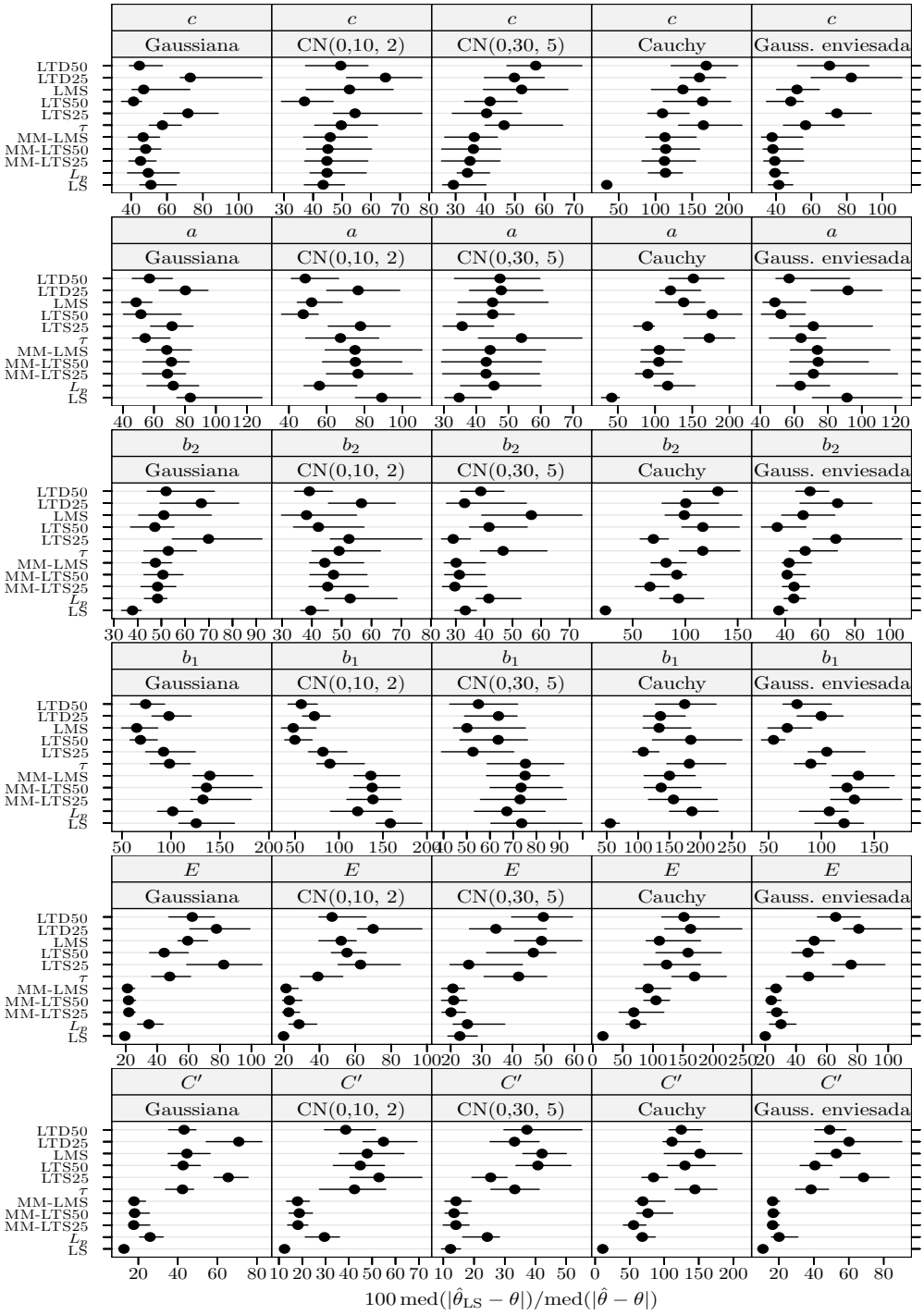


Figura 4.29 Lixiviação de minério manganífero: medida da eficiência dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 3, 15, 17, 18, 20, 21, 29, 33, 37, 43, 51, e 60.

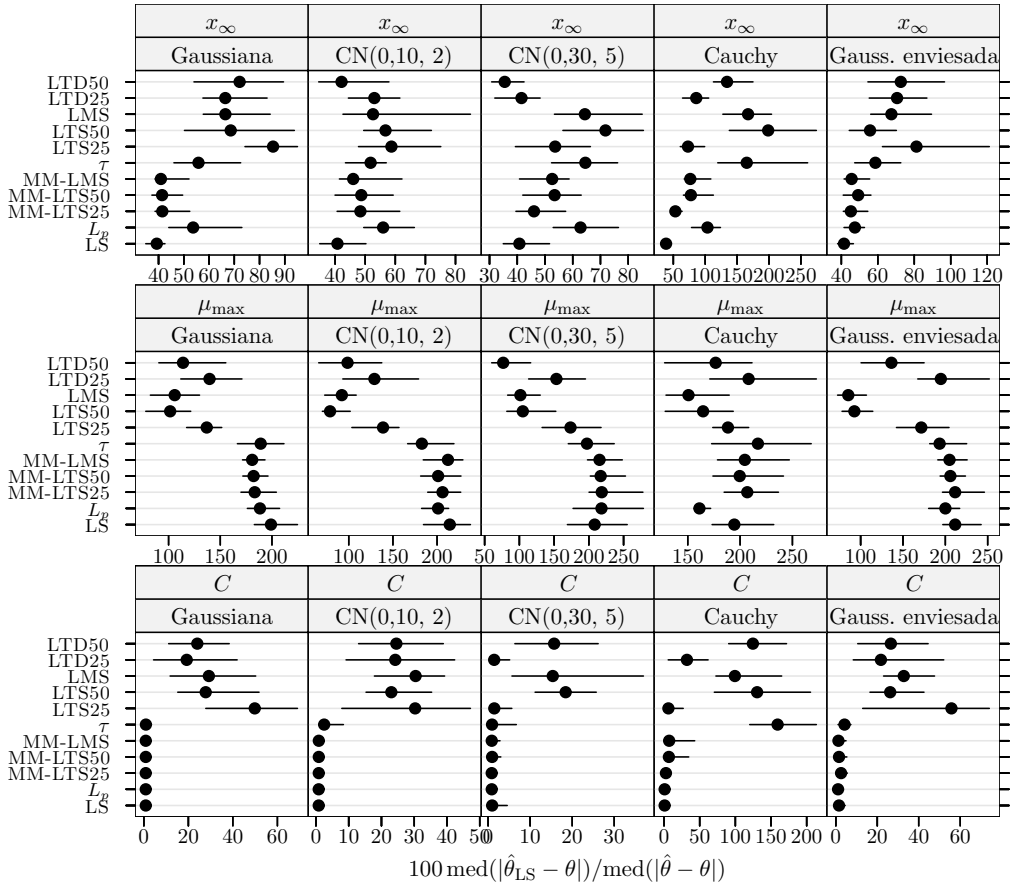


Figura 4.30 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: medida da eficiência dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 5, 11, e 12.

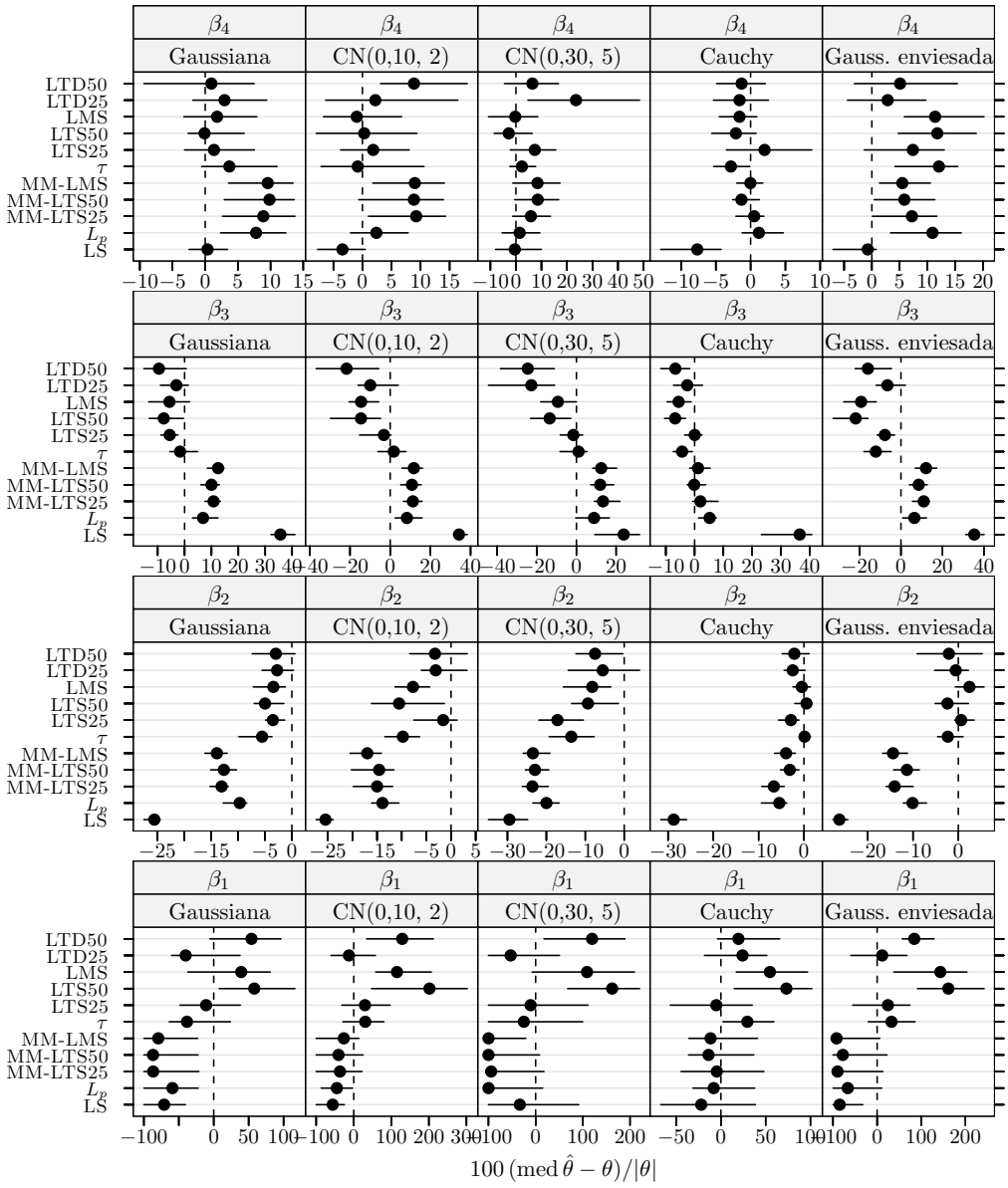


Figura 4.31 Isomerização do n-pentano: índice de enviesamento robustificado dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 11, 13, 14, 15, e 21.

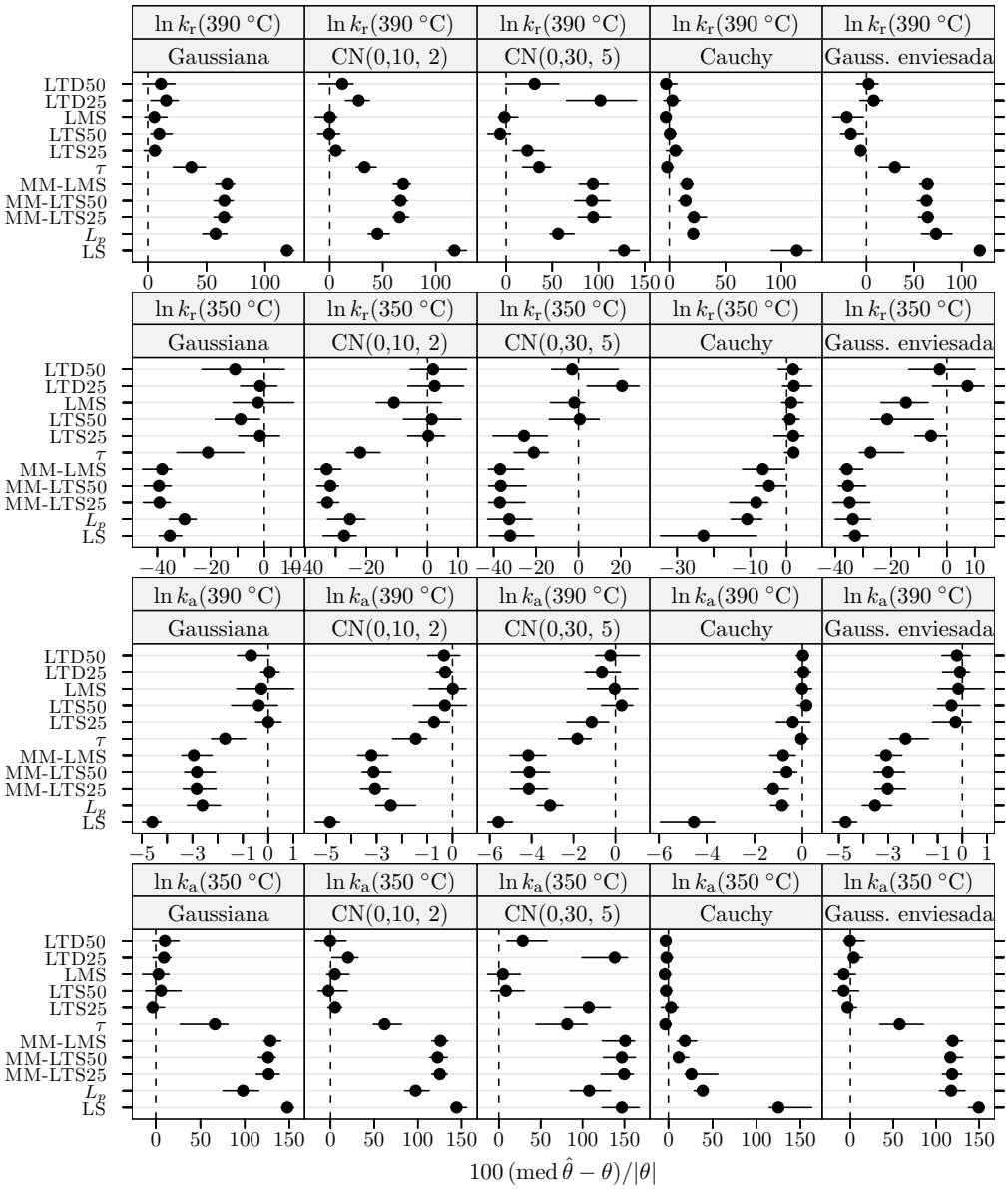


Figura 4.32 Oxidação do propeno: índice de enviesamento robustificado dos estimadores para dados simulados com 20% de *outliers* e $\delta_R = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 5, 8, 17, 19, 24, 32, 33, 39, 45, 47, 49, 56, e 66.

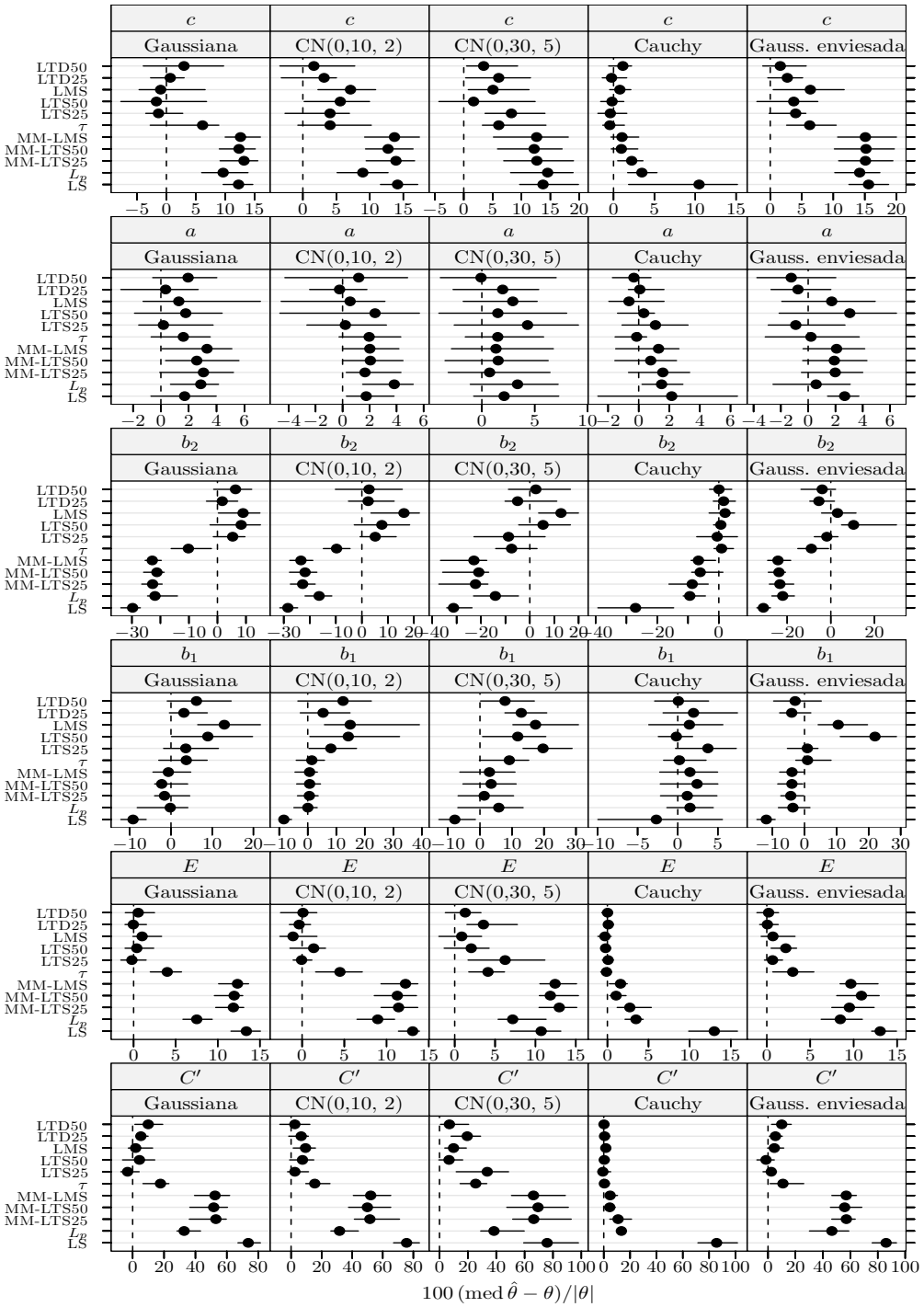


Figura 4.33 Lixiviação de minério manganífero: índice de enviesamento robustificado dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 3, 15, 17, 18, 20, 21, 29, 33, 37, 43, 51, e 60.

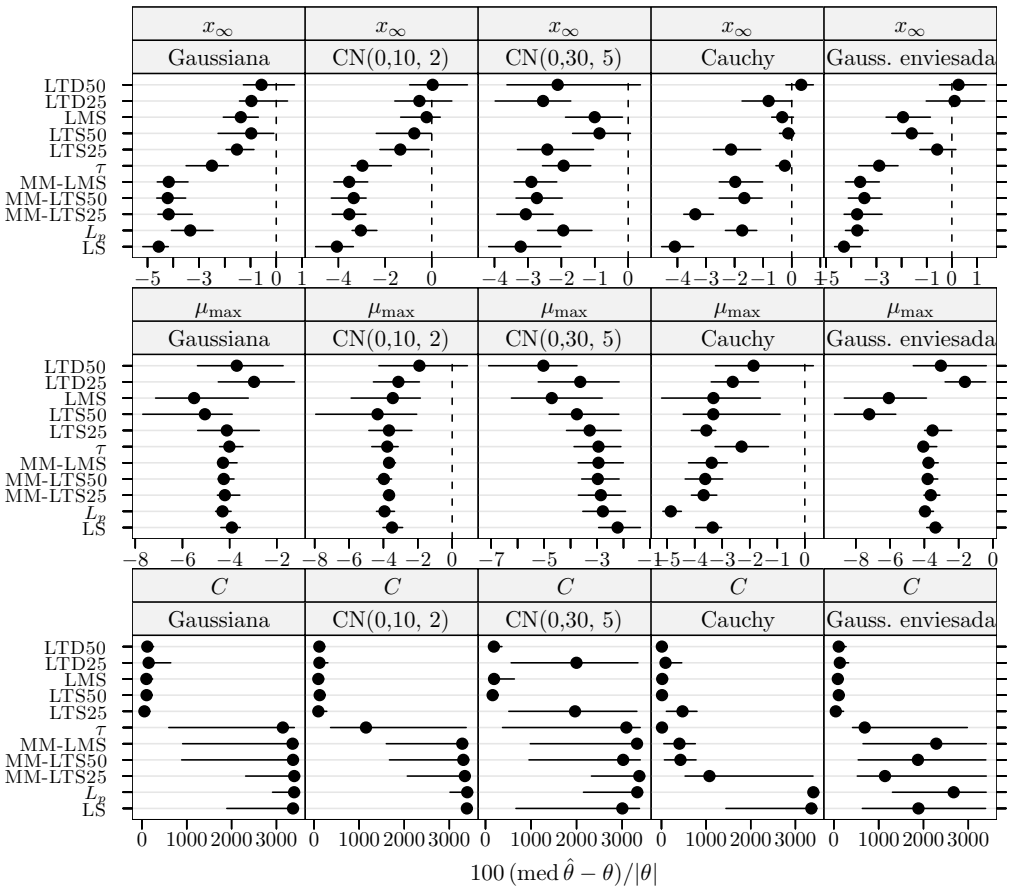


Figura 4.34 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: índice de enviesamento robustificado dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 5, 11, e 12.

De novo, verifica-se uma forte perda de eficiência dos estimadores do grupo I (embora notoriamente menos acentuado para o estimador τ) e LS, enquanto os estimadores do grupo II mantêm basicamente o nível de eficiência apresentado quando não existe contaminação por *outliers*. Este resultado era de esperar já que basicamente corresponde a um grau mais elevado do que se observa no cenário anterior. Apesar de as estimativas LS em geral apresentarem os valores de desempenho mais baixos, verificam-se curiosamente alguns resultados "anómalos" (*e.g.*, β_4 no modelo de isomerização do n-pentano, a no modelo de lixiviação de minério manganífero, $\mu_{\max}$ no modelo de crescimento de células MRC-5, todos sob as distribuições Gaussiana, CN(0,10, 2) e Gaussiana enviesada) onde têm o melhor desempenho. Repare-se que em alguns casos, nomeadamente E e C' no modelo de lixiviação de minério manganífero e $\ln k_{\mathrm{r}}(350\ ^\circ\mathrm{C})$ no modelo de oxidação do propeno, o grau de perda de eficiência é suficiente para inverter claramente o desempenho relativo entre os estimadores do grupo I e os estimadores MM e L_p. Nos casos restantes existe alguma evidência de uma tendência fraca quer na mesma direcção, quer na direcção oposta. De certo modo é como se este cenário configurasse um regime de transição. Globalmente, os estimadores LTS25 e τ apresentam tanto quanto possível o melhor desempenho.

Interessa ainda realçar a eficiência acrescida dos vários estimadores para $\mu_{\max}$ no modelo de crescimento de células MRC-5 em relação aos cenários anteriores, contrariamente ao que se poderia esperar.

Critério de enviesamento Contrariamente aos dois cenários anteriores, aqui observa-se um padrão para o enviesamento das estimativas: como regra geral as estimativas MM, L_p, e LS possuem valores de enviesamento superiores aos apresentados pelas estimativas do grupo II. Por seu lado, o estimador τ apresenta valores intermédios entre os dos grupos acima referidos. Em geral a magnitude do enviesamento permanece moderada. Contudo, é importante sublinhar que em relação aos estimadores MM, L_p e LS emergem aqui pela primeira vez casos (*e.g.*, $\ln k_{\mathrm{a}}(350\ ^\circ\mathrm{C})$) em que a magnitude (bastante elevada) do enviesamento retira significado prático às respectivas estimativas. Note-se, em particular, que todos os estimadores "rompem" para o parâmetro C do modelo de crescimento de células animais em *microcarriers*, embora os estimadores LMS, LTS e LTD sejam claramente mais resistentes.

Por fim, pela análise das figuras, verifica-se que o impacto dos *outliers* parece ser dominante relativamente às distribuições consideradas para o erro, já que não se observam características particularmente distintas entre os resultados associados às várias distribuições.

4.8.3 Caso de forte contaminação e perturbação (15% de outliers, $\delta_{\mathsf{R}} = 10$)

Como pode observar-se nas figuras 4.35 a 4.42 nas páginas 94–101, o padrão dos resultados segue de perto o do cenário de contaminação moderada que vimos anteriormente.

Assinala-se, em particular, a franca deterioração da eficiência do estimador LS quer em relação ao cenário sem *outliers* quer em comparação com o de todos os estimadores robustos no caso presente. Isto acontece também para o estimador L_p, embora com um grau consideravelmente mais baixo.

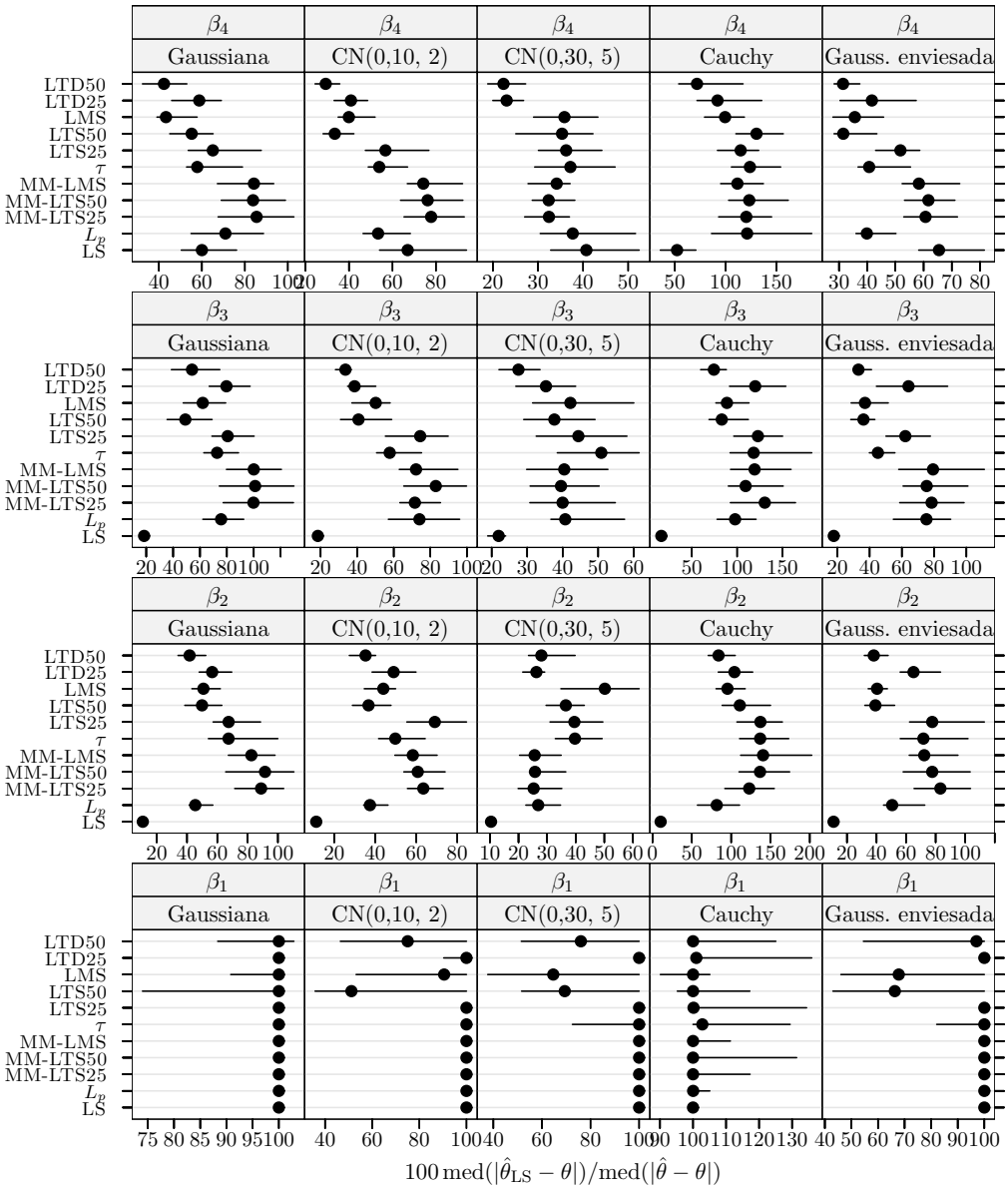


Figura 4.35 Isomerização do n-pentano: medida da eficiência dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 11, 13, 15, e 21.

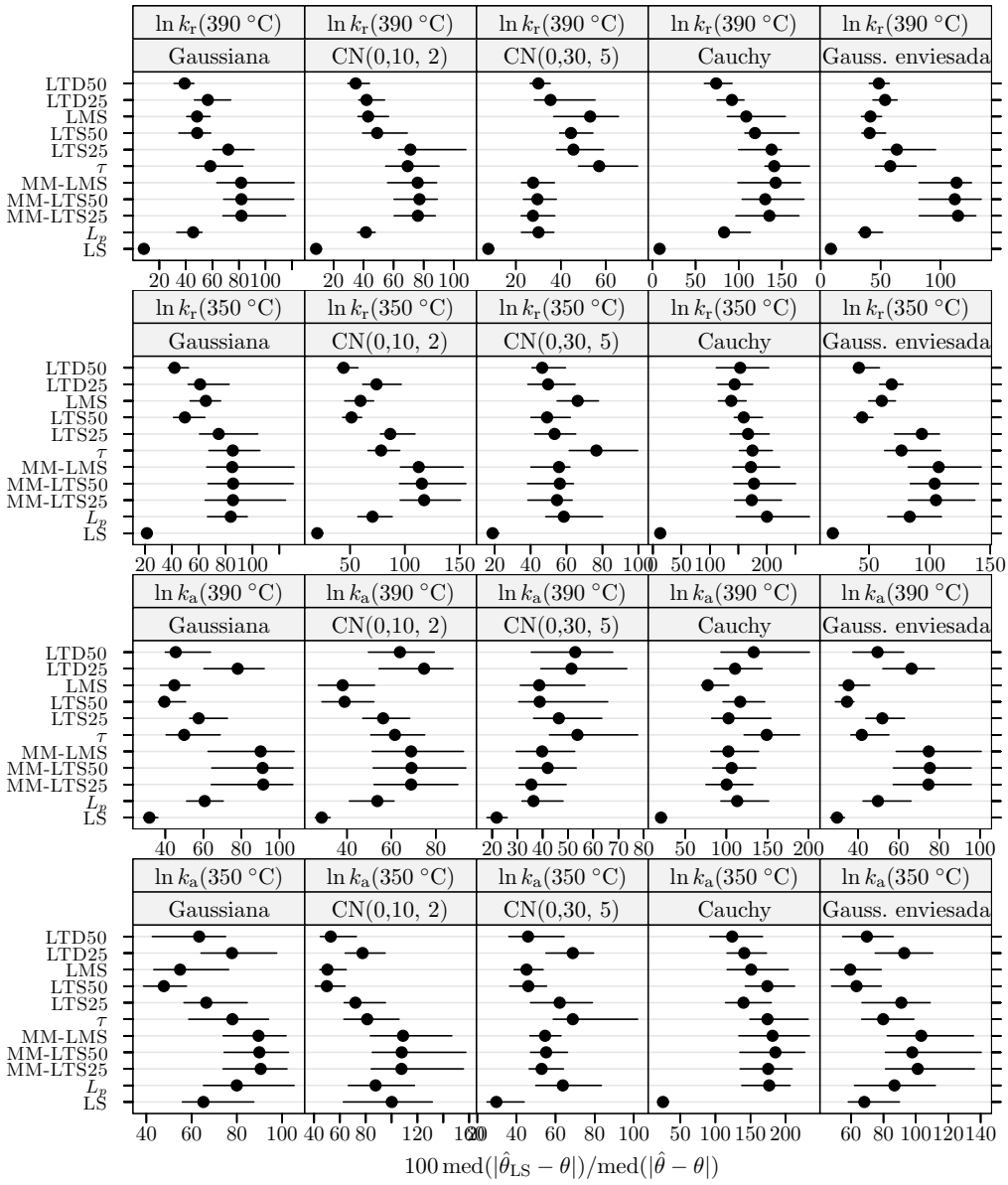


Figura 4.36 Oxidação do propeno: medida da eficiência dos estimadores para dados simulados com 15% de *outliers* e $\delta_R = 10$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 8, 24, 32, 33, 39, 45, 47, 49, 56, e 66.

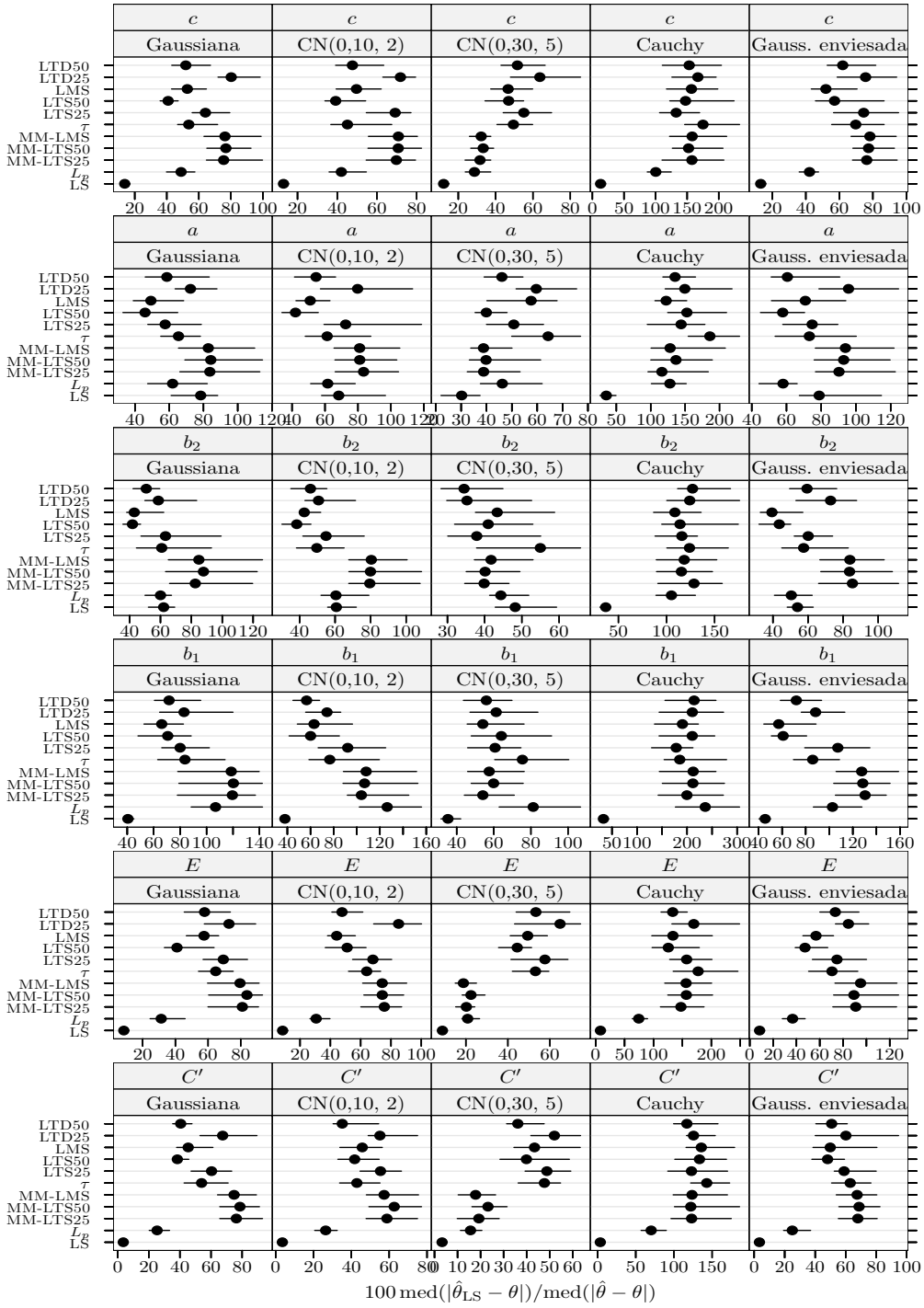


Figura 4.37 Lixiviação de minério manganífero: medida da eficiência dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 3, 15, 17, 18, 20, 21, 29, 33, e 60.

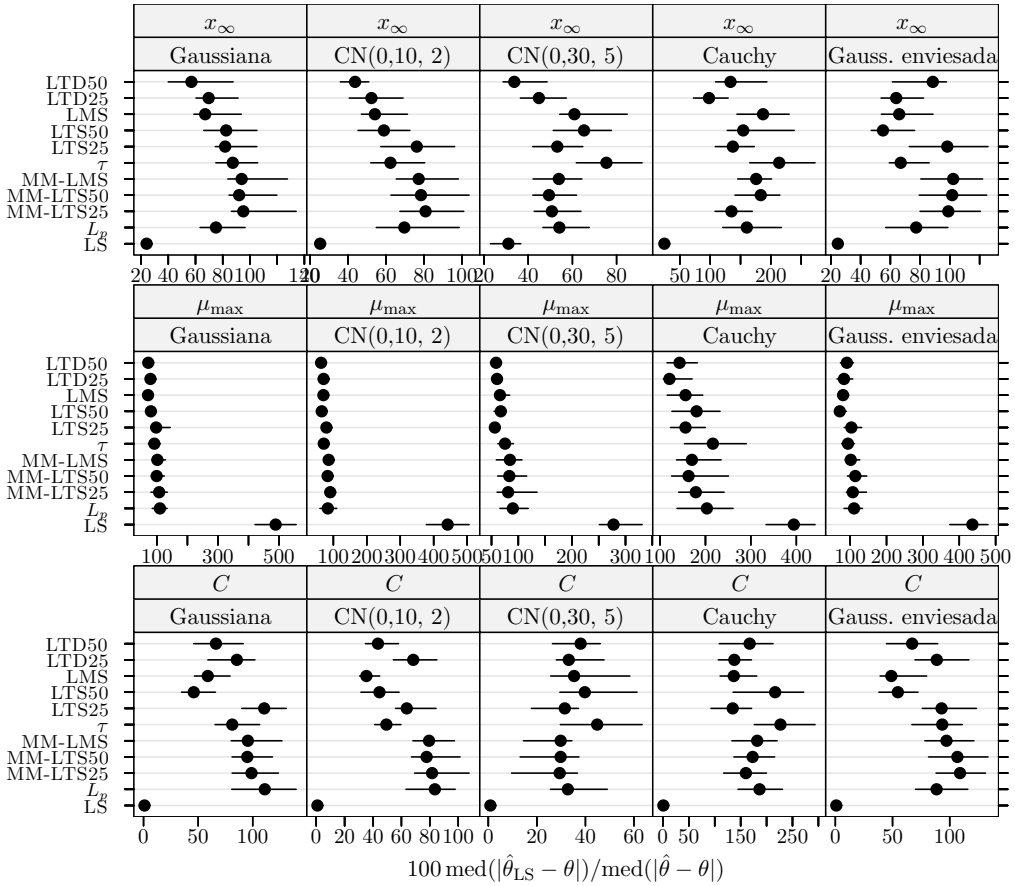


Figura 4.38 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: medida da eficiência dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 5, e 11.

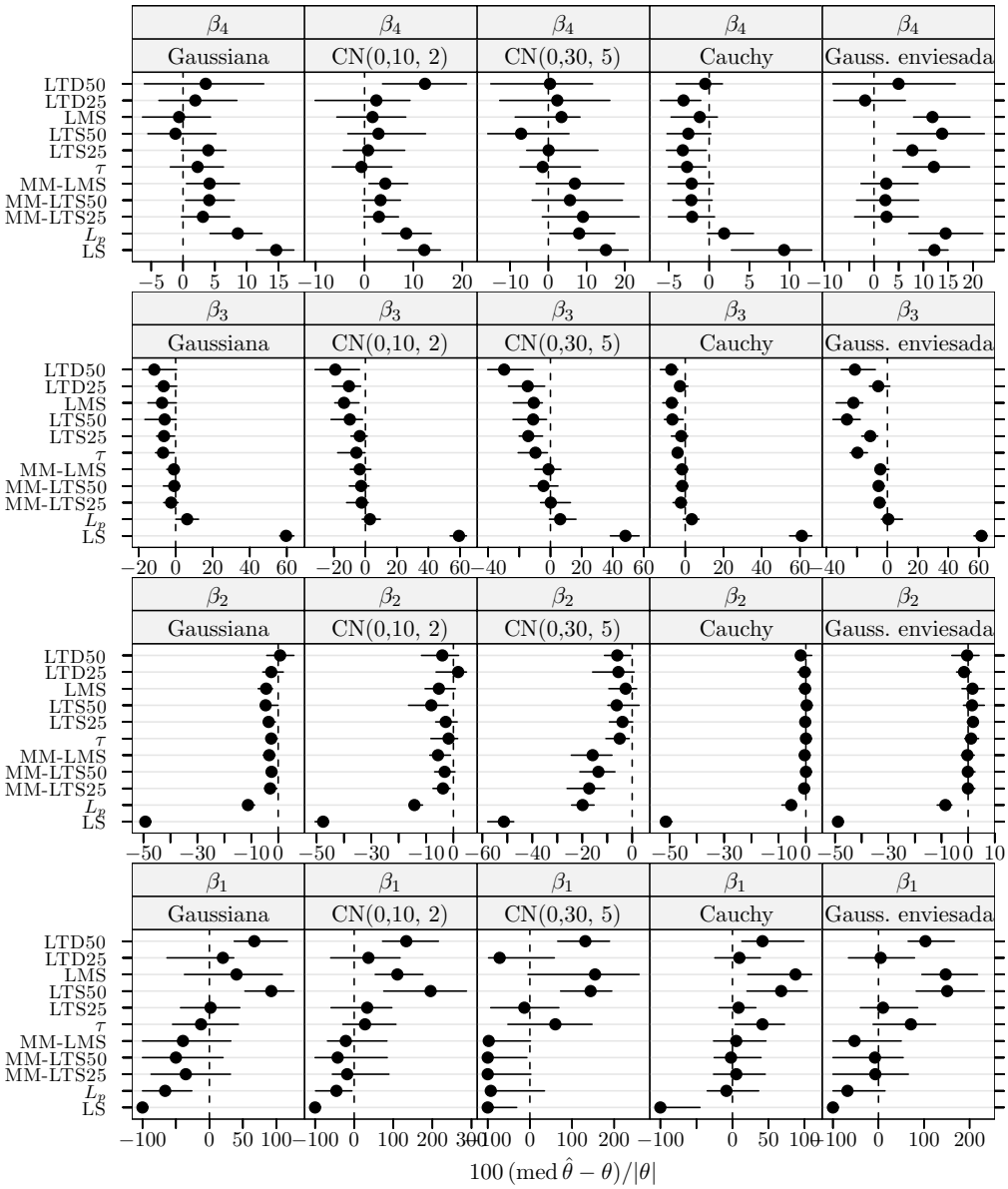


Figura 4.39 Isomerização do n-pentano: índice de enviesamento robustificado dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 11, 13, 15, e 21.

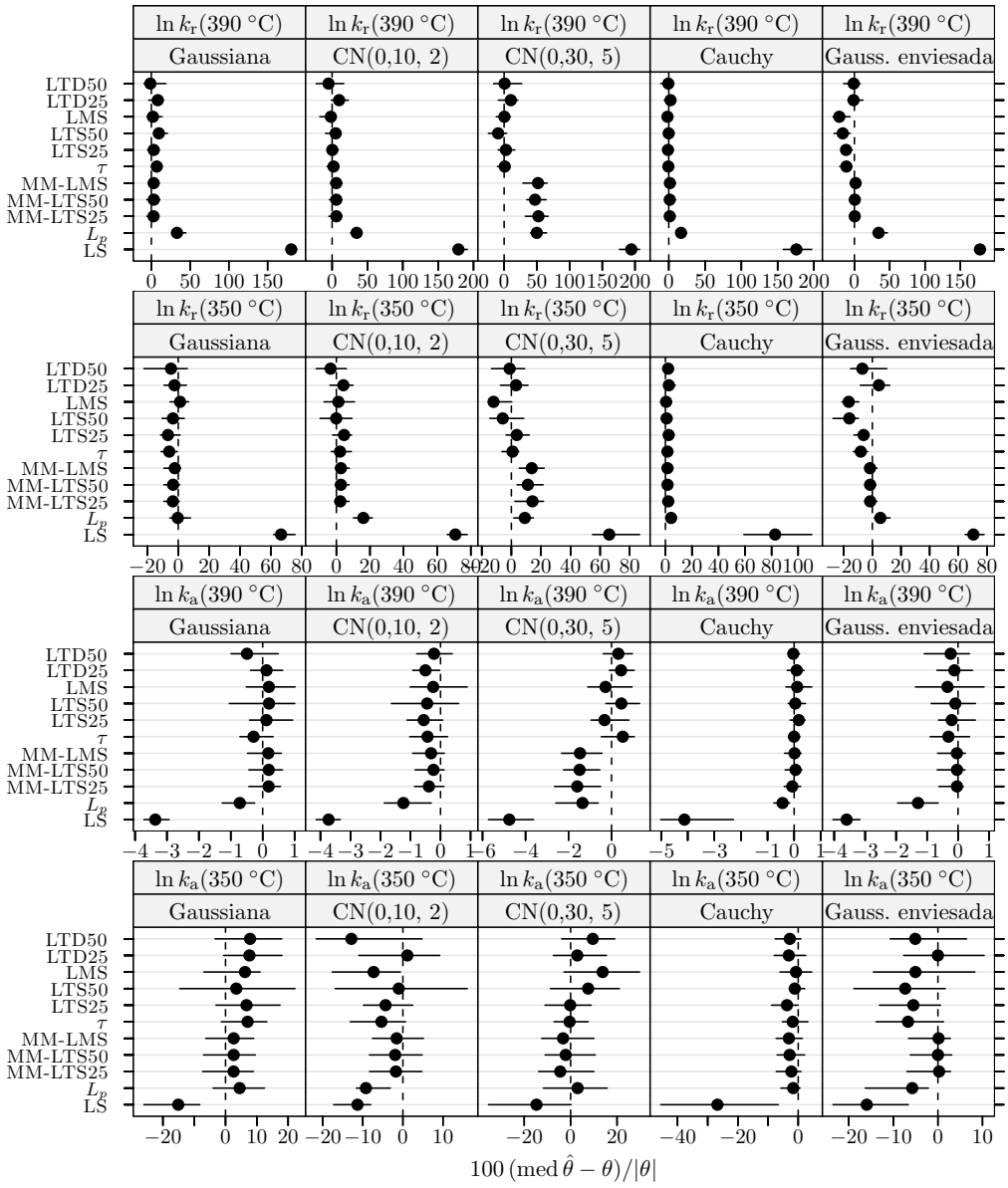


Figura 4.40 Oxidação do propeno: índice de enviesamento robustificado dos estimadores para dados simulados com 15% de *outliers* e $\delta_R = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 8, 24, 32, 33, 39, 45, 47, 49, 56, e 66.

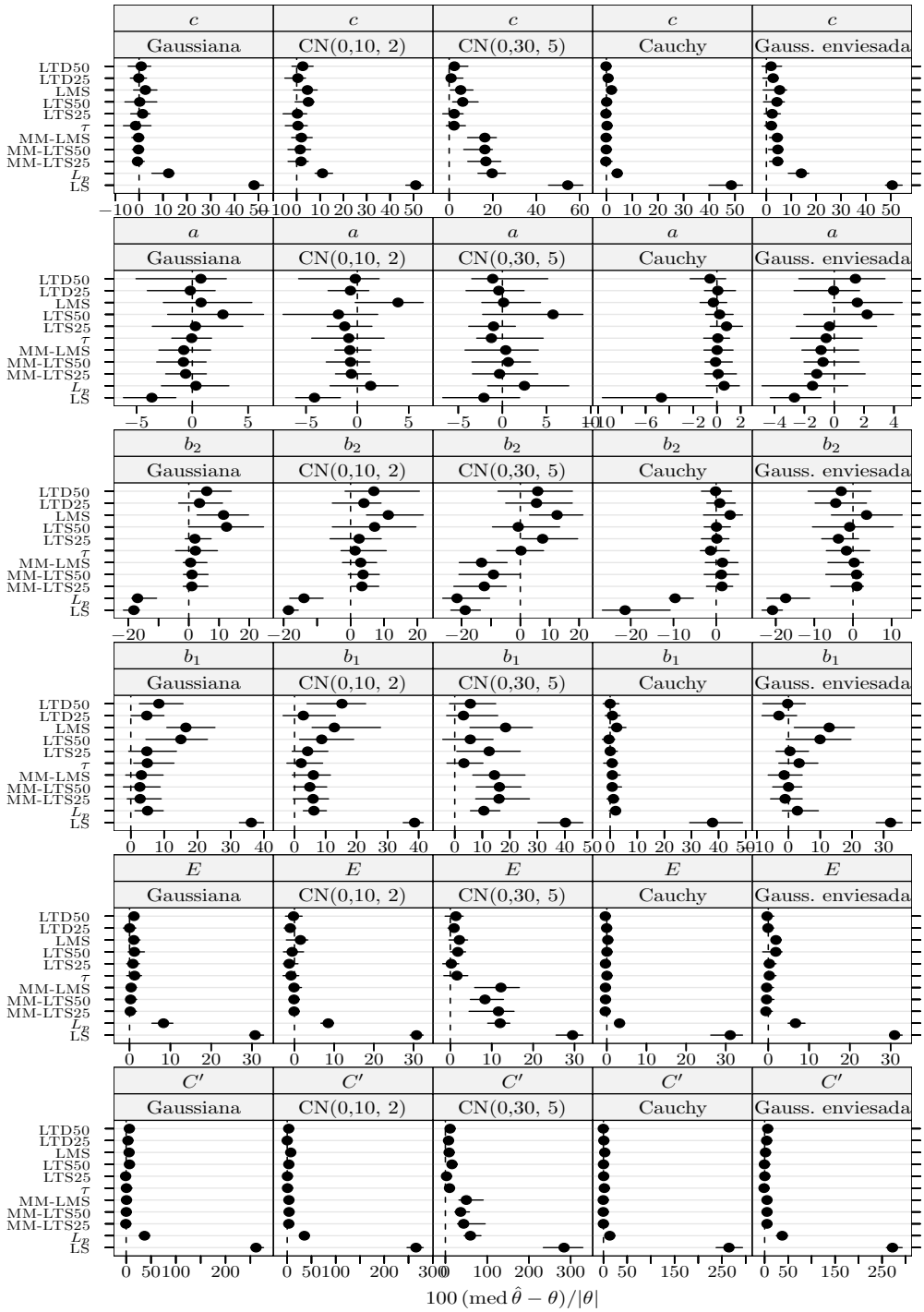


Figura 4.41 Lixiviação de minério manganífero: índice de enviesamento robustificado dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 3, 15, 17, 18, 20, 21, 29, 33, e 60.

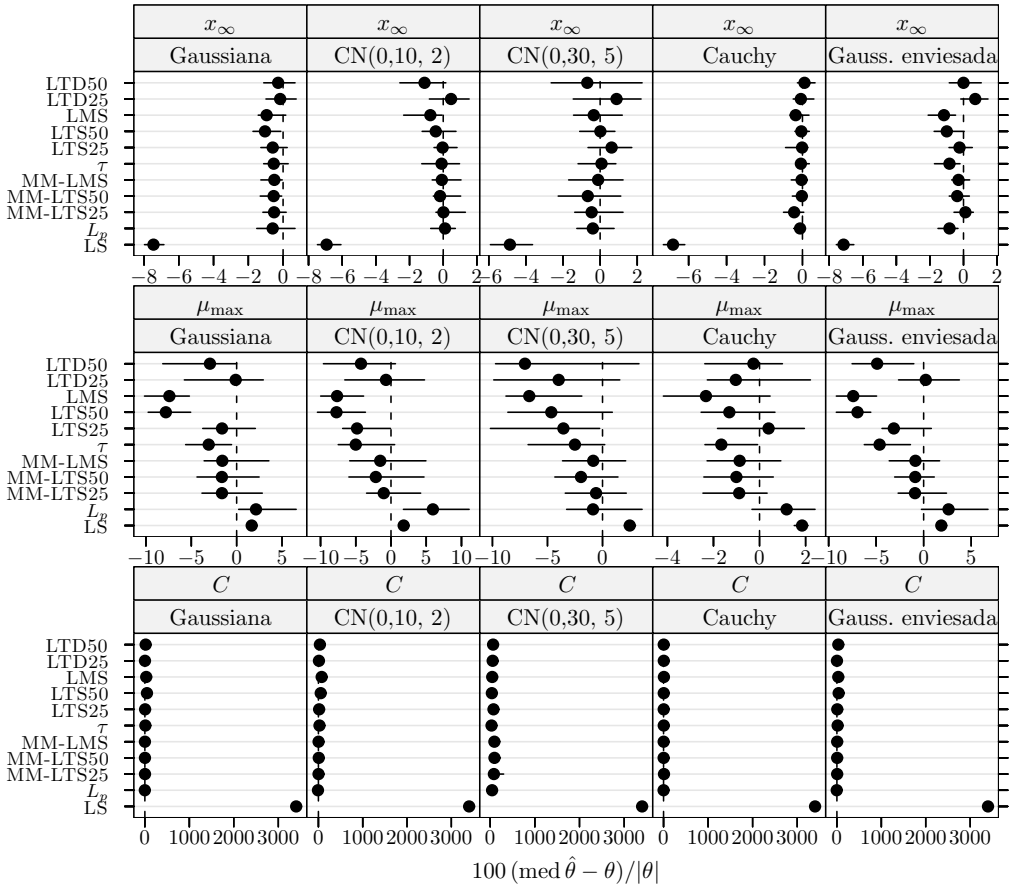


Figura 4.42 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: índice de enviesamento robustificado dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 5, e 11.

Constata-se ainda que as estimativas LS apresentam frequentemente valores substancialmente elevados de enviesamento. É o que se passa com os parâmetros $\ln k_{\mathrm{r}}(390\ ^{\circ}\mathrm{C})$ no modelo de oxidação do propeno, C' no modelo de lixiviação de minério manganífero, e C no modelo de crescimento de células MRC-5, onde o presente cenário conduz à rotura das estimativas LS obtidas.

Tal como visto no cenário de contaminação severa com este último modelo, observa-se também aqui uma melhoria da eficiência dos diversos estimadores para o parâmetro $\mu_{\max}$ em comparação com situações de grau baixo (ou nulo) de contaminação por *outliers*.

4.8.4 Caso limite (30% de outliers, $\delta_{\mathsf{R}} = 10$)

Conforme foi referido anteriormente, este cenário pretende representar circunstâncias extremas de contaminação.

Analisando agora as figuras 4.43 a 4.46 nas páginas 103–106 percebe-se a inversão dos níveis de desempenho dos grupos de estimadores robustos I e II, à excepção dos estimadores τ e LTS25. Concretamente, na maioria dos casos verifica-se a superioridade dos estimadores LMS, LTS50, e LTD sobre os estimadores MM e L_p. Ainda a este propósito, as variantes MM-LTS50 e MM-LMS apresentam de um modo consistente valores para a eficiência superiores à variante MM-LTS25 e ao estimador L_p.

Contrariamente aos outros estimadores do grupo I, em muitas situações o estimador τ apresenta valores para a eficiência muito próximos dos melhores. Por outro lado, tendo em conta que a proporção de *outliers* excede a fracção de aparamento do estimador LTS25, seria de esperar o desempenho bastante pobre, aliás como de facto se verifica. É interessante notar, no entanto, que este efeito não se verifica necessariamente em todos os parâmetros de um dado modelo como o revela o caso da isomerização do n-pentano (β_3) e da oxidação do propeno ($\ln k_{\mathrm{r}}(390\ ^{\circ}\mathrm{C})$ e $\ln k_{\mathrm{a}}(390\ ^{\circ}\mathrm{C})$).

Assim, pode-se pensar que a discrepância da variante MM-LTS25 em comparação com as restantes é devida à qualidade pobre de algumas das estimativas LTS25 necessárias à construção do estimador.

Curiosamente, repare-se no contraponto entre o quadro acabado de descrever e os resultados obtidos para o parâmetro $\mu_{\max}$ no modelo de crescimento de células MRC-5, em que os estimadores LTS25, MM (todas as variantes), e L_p apresentam claramente o melhor desempenho. Neste caso, ainda na perspectiva de "resultados anómalos" assinala-se, de novo, a melhoria da eficiência (com valores bastante elevados para os estimadores antecedentes) de todos os estimadores em comparação com situações de grau baixo (ou nulo) de contaminação por *outliers*.

Globalmente, o estimador LS exibe o pior desempenho.

Relativamente ao enviesamento, os estimadores LS, MM, e L_p apresentam em geral valores elevados ou bastante elevados e frequentemente sofrem rotura.

4.9 Comentários finais

Na presente secção vão-se descrever esquematicamente os principais aspectos do comportamento dos vários estimadores no contexto geral dos diversos cenários estudados.

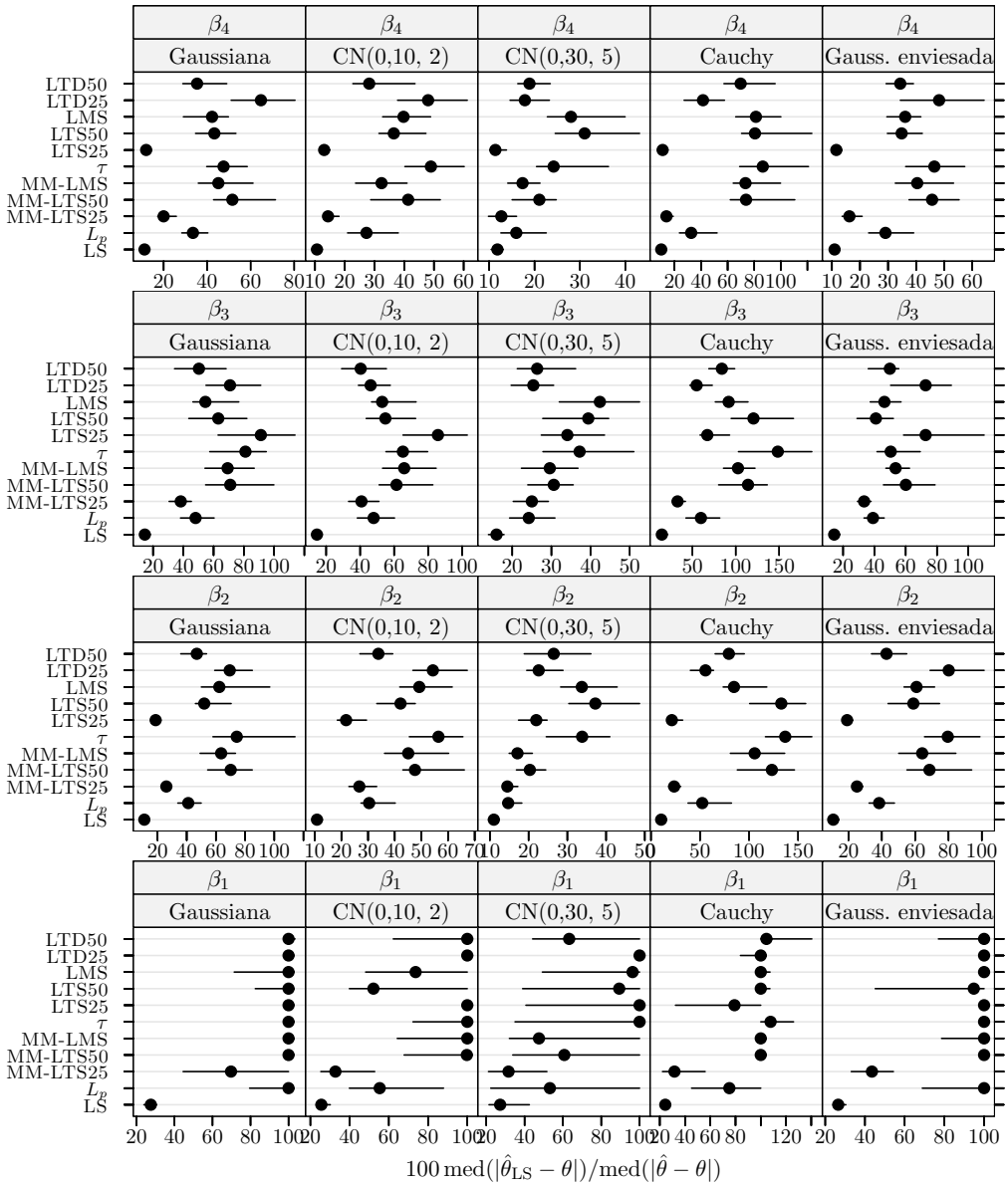


Figura 4.43 Isomerização do *n*-pentano: medida da eficiência dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 7, 11, 12, 13, 14, 15, e 21.

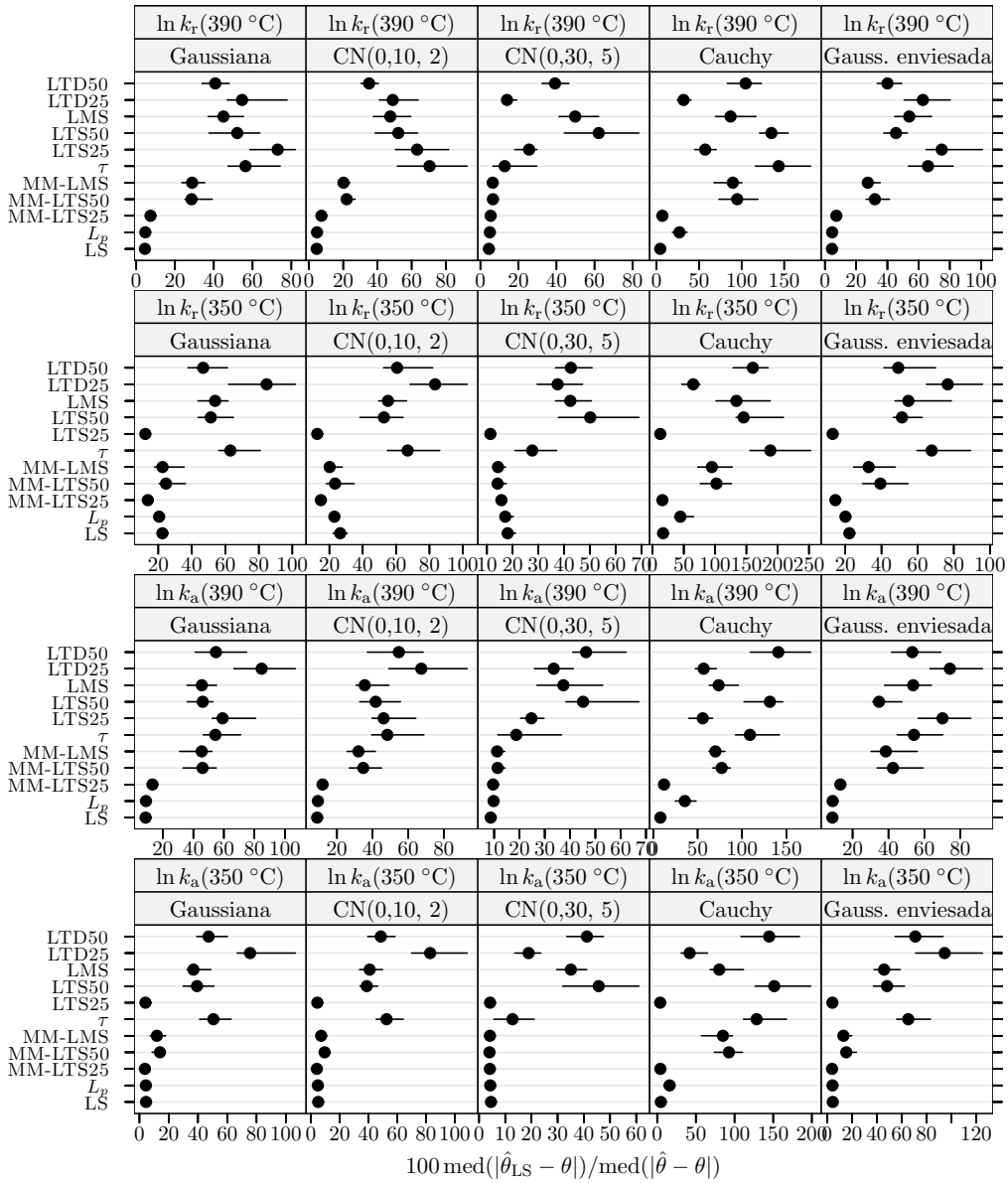


Figura 4.44 Oxidação do propeno: medida da eficiência dos estimadores para dados simulados com 30% de *outliers* e $\delta_R = 10$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 5, 8, 17, 18, 19, 24, 25, 27, 31, 32, 33, 36, 39, 42, 45, 47, 49, 52, 56, e 66.

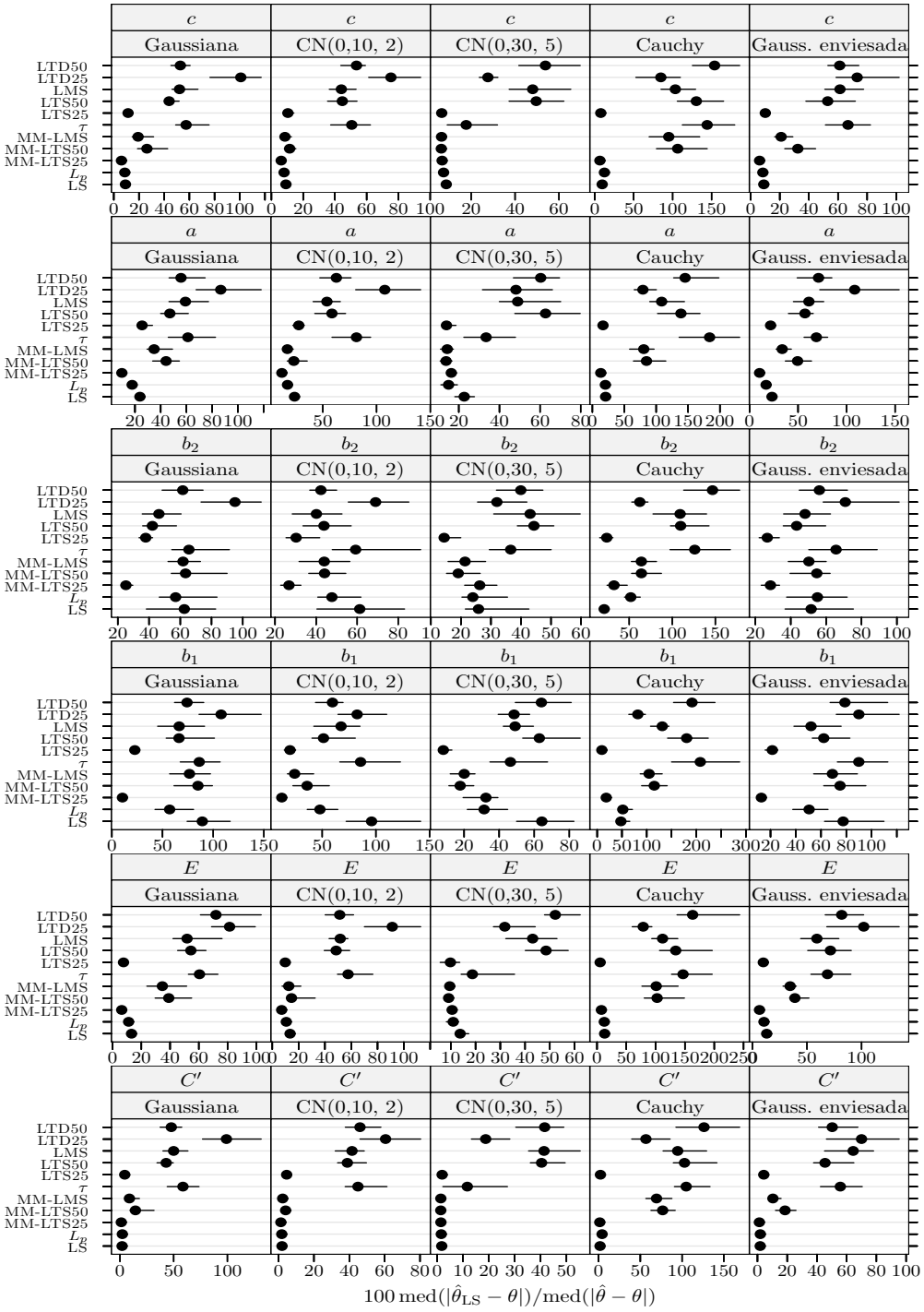


Figura 4.45 Lixiviação de minério manganífero: medida da eficiência dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 3, 6, 15, 16, 17, 18, 20, 21, 27, 29, 31, 33, 34, 37, 43, 51, 58, e 60.

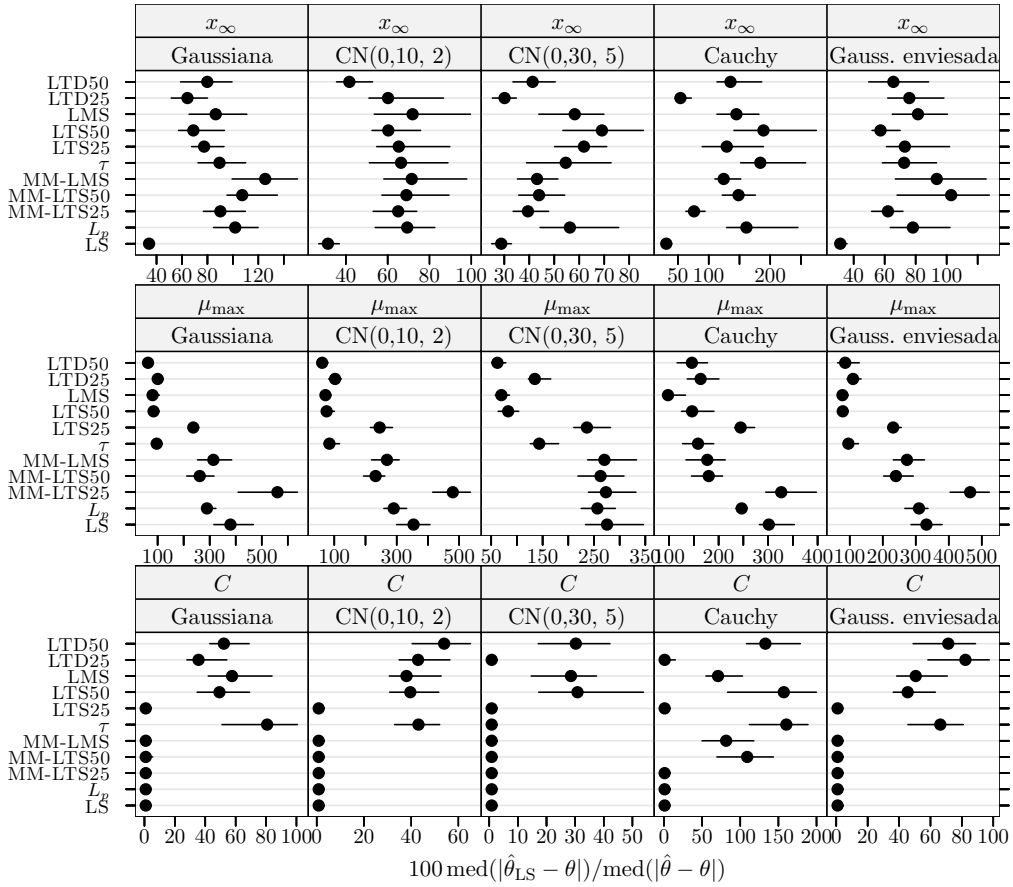


Figura 4.46 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: medida da eficiência dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador dos mínimos quadrados ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 5, 11, 12, e 14.

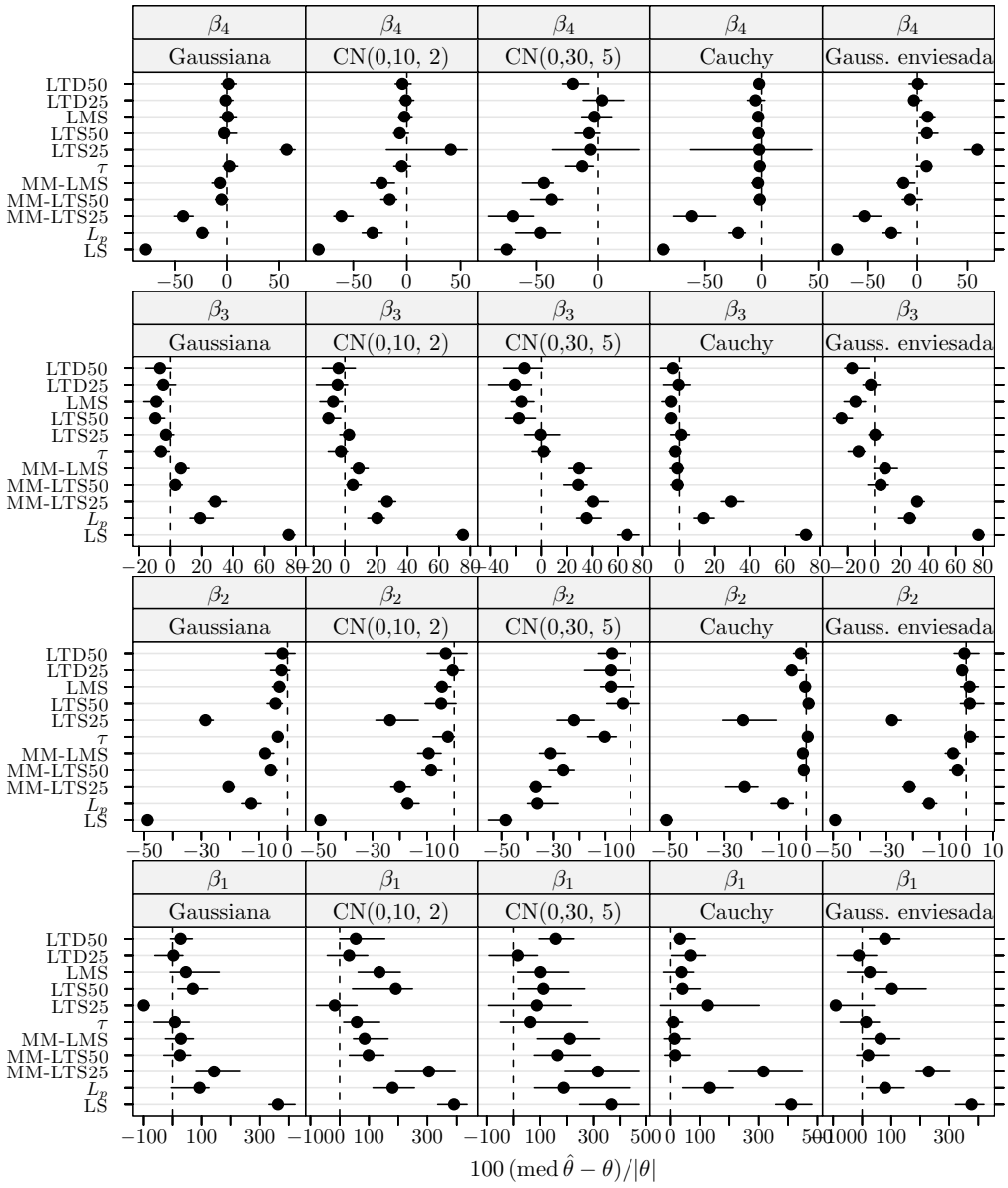


Figura 4.47 Isomerização do n-pentano: índice de enviesamento robustificado dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 7, 11, 12, 13, 14, 15, e 21.

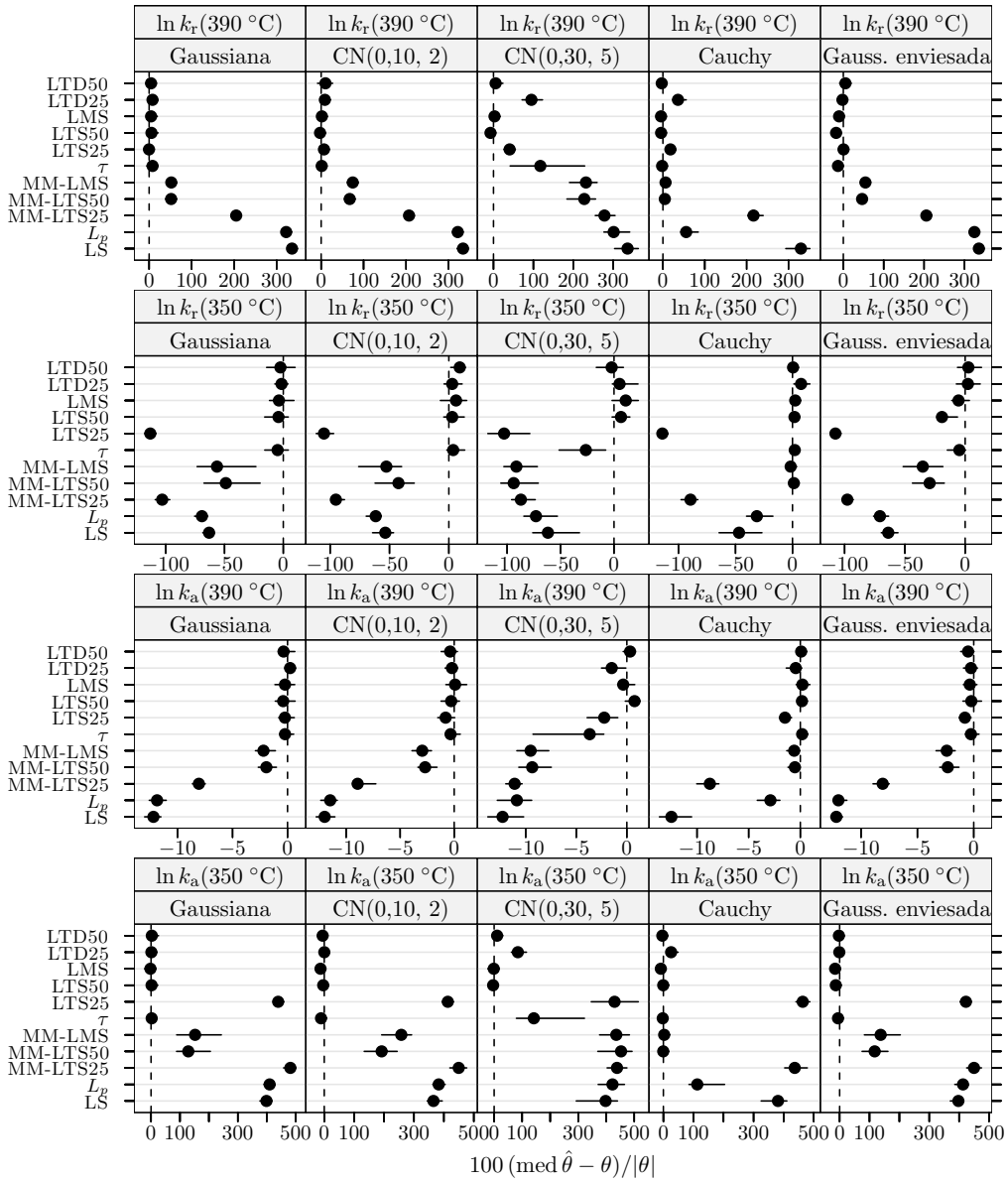


Figura 4.48 Oxidação do propeno: índice de enviesamento robustificado dos estimadores para dados simulados com 30% de *outliers* e $\delta_R = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 5, 8, 17, 18, 19, 24, 25, 27, 31, 32, 33, 36, 39, 42, 45, 47, 49, 52, 56, e 66.

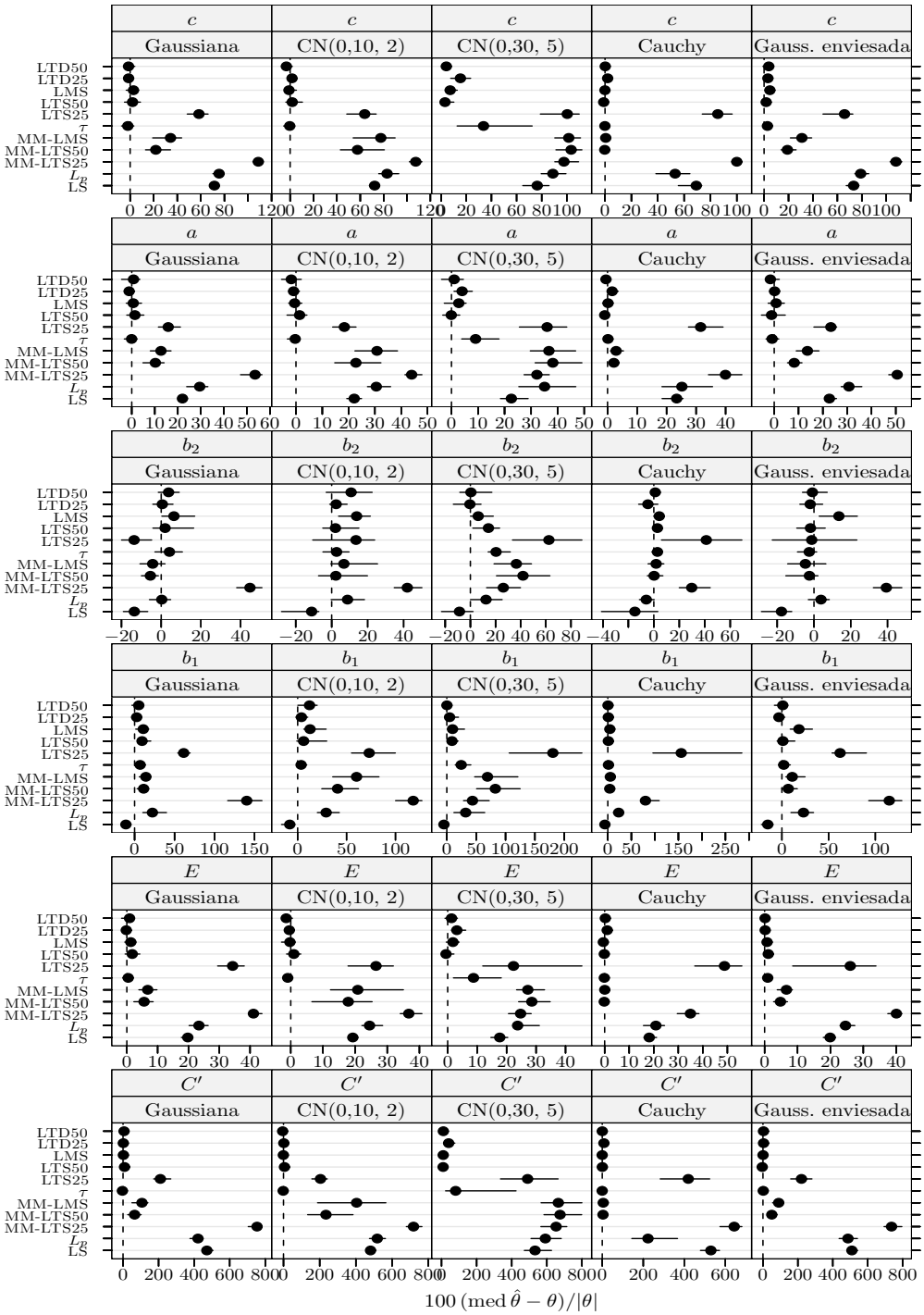


Figura 4.49 Lixiviação de minério manganífero: índice de enviesamento robustificado dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 3, 6, 15, 16, 17, 18, 20, 21, 27, 29, 31, 33, 34, 37, 43, 51, 58, e 60.

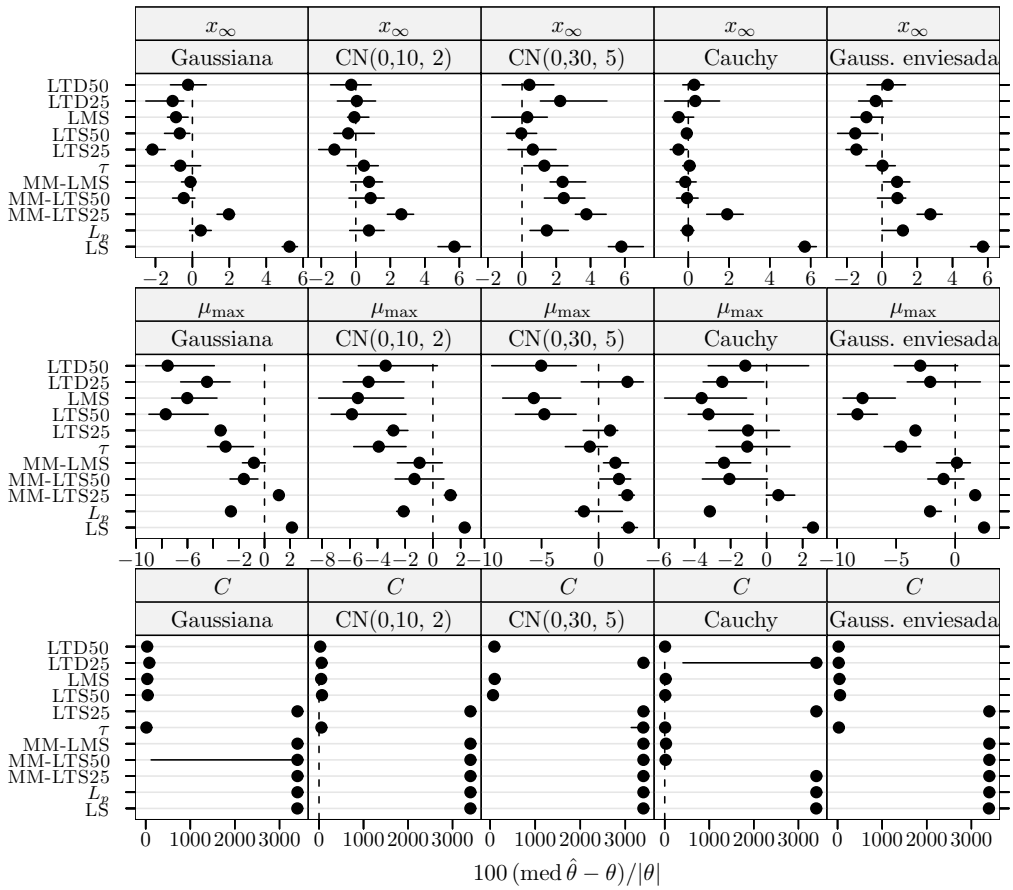


Figura 4.50 Crescimento de células MRC-5 em *microcarriers* Cytodex 1: índice de enviesamento robustificado dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 5, 11, 12, e 14.

1. Na ausência de *outliers* o estimador LS pode acomodar desvios suaves à distribuição Gaussiana sem degradação significativa do seu desempenho.

2. O estimador L_p tem um desempenho satisfatório quer na ausência de *outliers* quer com contaminação desde que, nesta último situação, o afastamento dos *outliers* no espaço-y seja pequeno e o nível de contaminação não seja extremo.

3. Em geral, os estimadores LMS e LTS50 têm um desempenho bastante insatisfatório. A razão disto resulta do efeito adverso — e particularmente significativo em pequenas amostras — de basicamente ajustarem apenas metade das observações. No entanto, sem surpresa, apresentam o melhor desempenho em situações limite de contaminação.

4. O estimador LTS25 apresenta um desempenho bastante satisfatório sob distribuições significativamente desviadas da Gaussiana, em situações de contaminação por *outliers*. Note-se que nestas circunstâncias o desempenho é semelhante ao do estimador τ. Como seria de esperar, se o nível de contaminação ultrapassa o valor estabelecido para o grau de aparamento, verifica-se uma deterioração substancial do desempenho.

5. Em princípio, as variantes do estimador MM com estimativas de elevado ponto de rotura fornecidas pelos estimadores LMS e LTS apresentam um desempenho essencialmente idêntico. Contudo, tendo presente os resultados do cenário com *outliers* e nível de contaminação extremo, pode-se conjecturar que situações em que o nível de contaminação ultrapasse o grau de aparamento usado para o estimador LTS conduzem a uma degradação significativa de desempenho. O desempenho é em geral semelhante ao dos estimadores LTS25 e τ; há pequenos indícios de uma ligeira superioridade sobre aqueles em situações de ausência de *outliers* ou de contaminação moderada.

6. De entre os estimadores em competição, o estimador τ é aquele com o domínio de aplicação mais largo porquanto mostra um desempenho bastante bom (frequentemente o melhor) em todos os cenários estudados.

7. O desempenho do estimador LTD é normalmente pobre e constitui mesmo uma desilusão no caso de erro assimétrico, onde, contrariamente ao que seria de esperar, não revela vantagens em comparação com os restantes estimadores robustos cuja construção suporta apenas distribuições simétricas para o erro aleatório. Porém, mostra um desempenho satisfatório em situações limite de contaminação.

Capítulo 5

Aplicações da simulação de Monte Carlo a modelos com resposta multivariada

5.1 Hidrogenação do tolueno

5.1.1 Descrição genérica do problema

Belohlav *et al.* (1997) relatam um estudo de determinação experimental da cinética da hidrogenação do tolueno. As experiências foram executadas à pressão e temperatura ambiente num reactor agitado semidescontínuo usando um catalisador comercial 5% Ru-act. A reacção decorre a temperatura constante. O esquema cinético proposto é o seguinte

$$\mathrm{A} \underset{k_{\mathrm{D},1}}{\overset{k_{\mathrm{H},1}}{\rightleftharpoons}} \mathrm{B} \xrightarrow{k_2} \mathrm{C}$$

onde A denota tolueno, B 1-metilciclo-hexeno, e C metilciclo-hexano, k_{H} representa as constantes de velocidade de hidrogenação (tempo^{-1}), k_{D} as constantes de velocidade de disproporcionação (tempo^{-1}), e $k_2 = k_{\mathrm{H},2} + k_{\mathrm{D},2}$.

As equações das velocidades de reacção correspondentes a este mecanismo são

$$\frac{\mathrm{d}c_{\mathrm{A}}}{\mathrm{d}t} = -k_{\mathrm{H},1}\theta_{\mathrm{A}} + k_{\mathrm{D},1}\theta_{\mathrm{B}} \tag{5.1a}$$

$$\frac{\mathrm{d}c_{\mathrm{B}}}{\mathrm{d}t} = k_{\mathrm{H},1}\theta_{\mathrm{A}} - (k_{\mathrm{D},1} + k_2)\theta_{\mathrm{B}} \tag{5.1b}$$

$$\frac{\mathrm{d}c_{\mathrm{C}}}{\mathrm{d}t} = k_2\theta_{\mathrm{B}} \tag{5.1c}$$

com as condições iniciais $c_{\mathrm{A}} = 1$, $c_{\mathrm{B}} = 0$, e $c_{\mathrm{C}} = 0$ para $t = 0$, onde c denota as concentrações normalizadas das várias espécies e t é o tempo. A fracção da superfície coberta pela espécie A, θ_{A}, e a fracção coberta por moléculas B, θ_{B}, são dadas pelas expressões

$$\theta_{\mathrm{A}} = \frac{K_{\mathrm{A}}^{\mathrm{rel}} c_{\mathrm{A}}}{K_{\mathrm{A}}^{\mathrm{rel}} c_{\mathrm{A}} + c_{\mathrm{B}} + K_{\mathrm{C}}^{\mathrm{rel}} c_{\mathrm{C}}} \tag{5.2a}$$

$$\theta_{\mathrm{B}} = \frac{c_{\mathrm{B}}}{K_{\mathrm{A}}^{\mathrm{rel}} c_{\mathrm{A}} + c_{\mathrm{B}} + K_{\mathrm{C}}^{\mathrm{rel}} c_{\mathrm{C}}} \tag{5.2b}$$

onde K^{rel} representa as constantes de adsorção relativas à constante de adsorção da espécie B, isto é, $K^{\mathrm{rel}} = K/K_{\mathrm{B}}$.

Tabela 5.1 Variação das concentrações de tolueno, 1-metilciclo-hexeno, e metilciclo-hexano ao longo do tempo

Obs.	t/min	c_{A}	c_{B}	c_{C}
1	15	0,695	0,312	0,001
2	30	0,492	0,430	0,080
3	45	0,276	0,575	0,151
4	60	0,225	0,570	0,195
5	75	0,163	0,575	0,224
6	90	0,134	0,533	0,330
7	120	0,064	0,462	0,471
8	180	0,056	0,362	0,580
9	240	0,041	0,211	0,747
10	320	0,031	0,146	0,822
11	360	0,022	0,080	0,898
12	380	0,021	0,070	0,909
13	400	0,019	0,073	0,908

Tabela 5.2 Hidrogenação do tolueno: parâmetros do algoritmo de optimização MDE usados na regressão dos dados experimentais

Espaço de procura $[10^{-80}, 10^{-80}, 10^{-80}, 10^{-80}, 10^{-80}] \leqslant [k_{\mathrm{H},1}, k_{\mathrm{D},1}, k_2, K_{\mathrm{A}}^{\mathrm{rel}}, K_{\mathrm{B}}^{\mathrm{rel}}] \leqslant [0{,}1, 1, 1, 100, 100]$					
N_{P}	F_{b}	C_{R}	ξ	Tolerância	Ponto incluído na população inicial
51	0,5	0,8	1,5	10^{-8}	$[0{,}023, 0{,}005, 0{,}011, 1{,}9, 1{,}8]$[a]

[a] Tabela 4 de Belohlav *et al.* (1997, p. 739).

Os dados experimentais encontram-se na tabela 5.1.

Na utilização da rotina de integração `LSODA`, estabeleceu-se 10^{-6} para as tolerâncias relativa (RTOL) e absoluta (ATOL).

5.1.2 Aplicação aos dados experimentais

Na tabela 5.2 encontram-se os parâmetros do método MDE usados no cálculo das diversas estimativas dos parâmetros do modelo (5.1) obtidas a partir dos dados experimentais, obtendo-se os valores apresentados na tabela 5.3 na próxima página. Relativamente aos erros de medição $\epsilon_k, k = (c_{\mathrm{A}}, c_{\mathrm{B}}, c_{\mathrm{C}})$, a matriz das estimativas das covariâncias, $\widehat{\boldsymbol{\Sigma}}$, obtida para o estimador ML (critério do determinante) e a correspondente matriz de correlações, $\widehat{\boldsymbol{\Psi}}$, são

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0{,}000234 & & \\ -0{,}000145 & 0{,}000514 & \\ -0{,}000101 & -0{,}000365 & 0{,}000598 \end{bmatrix}, \quad \widehat{\boldsymbol{\Psi}} = \begin{bmatrix} 1 & & \\ -0{,}417 & 1 & \\ -0{,}269 & -0{,}658 & 1 \end{bmatrix}.$$

Como se pode observar, qualquer das estimativas robustas pouco diferem daquelas correspondentes ao estimador ML, à excepção das estimativas MTL50 dos parâmetros

Tabela 5.3 Hidrogenação do tolueno: estimativas para os parâmetros do modelo

	Parâmetros				
Estimadores	$k_{\mathrm{H},1}/\mathrm{min}^{-1}$	$k_{\mathrm{D},1}/\mathrm{min}^{-1}$	k_2/min^{-1}	$K_{\mathrm{A}}^{\mathrm{rel}}$	$K_{\mathrm{B}}^{\mathrm{rel}}$
ML	0,023	0,0051	0,011	1,9	1,8
LAD	0,025	0,0046	0,0095	1,4	1,3
M	0,024	0,0035	0,0091	1,4	1,3
MML	0,022	0,019	0,015	4,2	2
MTL25	0,023	0,0075	0,011	2,2	1,7
MTL50	0,021	0,49	0,21	100	29

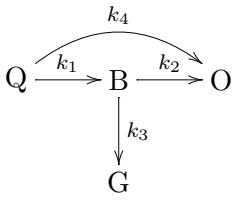
$k_{\mathrm{D},1}$, k_2, $K_{\mathrm{A}}^{\mathrm{rel}}$ e $K_{\mathrm{B}}^{\mathrm{rel}}$, que têm um afastamento bastante pronunciado.

5.2 Pirólise do xisto betuminoso

5.2.1 Descrição genérica do problema

O xisto betuminoso contém matéria orgânica insolúvel (querogeno) ligada à estrutura da rocha. A decomposição térmica desta substância por aquecimento sem reacção com o oxigénio (pirólise) produz óleo combustível, o qual pode substituir o petróleo bruto.

Há vários estudos sobre a cinética da pirólise de xisto betuminoso, nomeadamente os baseados em dados obtidos por Hubbard e Robinson (1950), reproduzidos na tabela 5.4 na página seguinte. Segundo o mecanismo proposto por Ziegel e Gorman (1980), em que as velocidades de reacção das diferentes etapas são de primeira ordem, pode-se escrever



onde Q denota querogeno, B betume, O óleo, e G outros produtos de decomposição que incluem resíduos orgânicos insolúveis (coque) e gases leves, e k refere-se às constantes de velocidade dos diferentes passos expressas (tempo^{-1}), que obedecem à lei de Arrhenius,

$$k = A \exp\Big(-\frac{E}{RT}\Big).$$

As leis cinéticas que descrevem esta sequência de passos são

$$\frac{\mathrm{d}c_{\mathrm{Q}}}{\mathrm{d}t} = -(k_1 + k_4)c_{\mathrm{Q}} \tag{5.3a}$$

$$\frac{\mathrm{d}c_{\mathrm{B}}}{\mathrm{d}t} = k_1 c_{\mathrm{Q}} - (k_2 + k_3)c_{\mathrm{B}} \tag{5.3b}$$

$$\frac{\mathrm{d}c_{\mathrm{O}}}{\mathrm{d}t} = k_4 c_{\mathrm{Q}} + k_2 c_{\mathrm{B}}, \tag{5.3c}$$

Tabela 5.4 Evolução temporal da fracção (expressa em percentagem da massa inicial de querogeno) do betume e óleo formados para várias temperaturas na pirólise de xisto betuminoso

$T = 673$ K				$T = 698$ K			
Obs.	t/min	Betume	Óleo	Obs.	t/min	Betume	Óleo
1	5	0,0	0,0	15	5,0	6,5	0,0
2	7	2,2	0,0	16	7,0	14,4	1,4
3	10	11,5	0,7	17	10,0	18,0	10,8
4	15	13,7	7,2	18	12,5	16,5	14,4
5	20	15,1	11,5	19	15,0	29,5	21,6
6	25	17,3	15,8	20	17,5	23,7	30,2
7	30	17,3	20,9	21	20,0	36,7	33,1
8	40	20,1	26,6	22	25,0	27,3	40,3
9	50	20,1	32,4	23	30,0	16,5	47,5
10	60	22,3	38,1	24	40,0	7,2	55,4
11	80	20,9	43,2	25	50,0	3,6	56,8
12	100	11,5	49,6	26	60,0	2,2	59,7
13	120	6,5	51,8				
14	150	3,6	54,7				
$T = 723$ K				$T = 748$ K			
27	5,0	8,6	0,0	39	3,0	0,7	0,0
28	7,5	15,8	2,9	40	4,5	17,3	2,9
29	8,0	25,9	16,5	41	5,0	23,0	17,3
30	9,0	25,2	24,4	42	5,5	24,4	20,9
31	10,0	26,6	29,5	43	6,0	23,0	25,9
32	11,0	33,8	35,2	44	6,5	33,1	29,5
33	12,5	25,9	39,5	45	7,0	31,6	33,8
34	15,0	20,1	45,3	46	8,0	20,9	45,3
35	17,5	12,9	43,1	47	9,0	10,1	53,2
36	17,5	9,3	54,6	48	10,0	4,3	58,2
37	20,0	3,6	59,7	49	12,5	0,7	57,5
38	20,0	2,2	53,9	50	15,0	0,7	61,1
$T = 773$ K				$T = 798$ K			
51	3,0	6,5	0,0	59	3,00	25,2	20,9
52	4,0	24,4	23,0	60	3,25	33,1	25,2
53	4,5	26,6	32,4	61	3,50	21,6	17,3
54	5,0	25,9	37,4	62	4,00	20,9	36,7
55	5,5	17,3	45,3	63	5,00	4,3	56,8
56	6,0	21,6	45,3	64	7,00	0,0	61,8
57	6,5	1,4	57,5				
58	10,0	0,0	60,4				

Fonte: Tabela A1.15 de Bates e Watts (1988, pp. 283 e 284).

com condições iniciais

$$c_{\mathrm{Q}}(t_0) = 1, \quad c_{\mathrm{B}}(t_0) = 0, \quad c_{\mathrm{O}}(t_0) = 0, \tag{5.3d}$$

onde c refere-se às concentrações das várias espécies, t é o tempo, e t_0 é um "tempo morto" que dá uma indicação sobre o intervalo de tempo que decorre até o xisto atingir a temperatura de iniciação do processo de decomposição térmica.

Da integração das equações (5.3) obtém-se

$$c_{\mathrm{Q}} = \mathrm{e}^{-\gamma\tau} \tag{5.4a}$$

$$c_{\mathrm{B}} = \frac{k_1}{\alpha}(\mathrm{e}^{-\beta\tau} - \mathrm{e}^{-\gamma\tau}) \tag{5.4b}$$

$$c_{\mathrm{O}} = \frac{k_1 k_2}{\alpha\beta}(1 - \mathrm{e}^{-\beta\tau}) + \frac{\alpha k_4 - k_1 k_2}{\alpha\gamma}(1 - \mathrm{e}^{-\gamma\tau}) \tag{5.4c}$$

onde $\alpha = k_1 + k_4 - k_2 - k_3$, $\beta = k_2 + k_3$, $\gamma = k_1 + k_4$, e $\tau = t - t_0$.

Por outro lado, Bates e Watts (1988, p. 194) sugerem utilizar uma relação linear entre t_0 e a quantidade

$$T_{\mathrm{inv}} = -\frac{1}{R}\Big(\frac{1}{T} - \frac{1}{723}\Big)$$

para descrever o efeito da temperatura sobre o "tempo morto". Aqui a constante dos gases é expressa em kJ/mol K.

Para reduzir as correlações entre as diferentes energias de activação e os correspondentes factores pré-exponenciais usou-se novamente a reparametrização exposta na secção 4.2.1 na página 48. Há pois dez parâmetros neste modelo: quatro logaritmos de constantes de velocidade a 673 K, mais quatro logaritmos de constantes de velocidade a 798 K, o "tempo morto" a 723 K, e o declive de t_0 em função de T_{inv}, b.

Na utilização do algoritmo MDE, estabeleceu-se 0,5 como limite superior do espaço de procura dos elementos da matriz de covariâncias entre respostas para as simulações de Monte Carlo dos estimadores M, MML, e MTL.

5.2.2 Aplicação aos dados experimentais

Os valores usados para os parâmetros do algoritmo MDE são: $N_{\mathrm{P}} = 81$, $F_{\mathrm{b}} = 0{,}5$, $C_{\mathrm{R}} = 0{,}8$, e $\xi = 1{,}5$; a tolerância usada para o critério de paragem foi 10^{-15}. A tabela 5.5 na página seguinte indica o espaço de procura e as estimativas incluidas na população inicial.

Na tabela 5.6 na próxima página encontram-se as estimativas dos parâmetros do modelo (5.3) dos diferentes estimadores resultantes do ajuste simultâneo dos dados das duas respostas medidas, óleo e betume, para todas as seis temperaturas. Na figura 5.1 na página 119 apresentam-se gráficos para os resultados referentes à tabela 5.6 na próxima página. Em geral, verifica-se que as diferentes estimativas não diferem muito entre si, Observa-se uma diferença (embora pequena) entre a estimativa MML e as restantes para os $\ln k_i(673\ \mathrm{K})$. Relativamente aos $\ln k_i(798\ \mathrm{K})$ observa-se uma diferença algo razoável entre as estimativas MML, MTL50, e MTL25, por uma parte, e M, LAD e ML, por outra parte.

Tabela 5.5 Pirólise do xisto betuminoso: dados para o algoritmo MDE

	Ponto incluído na	Espaço de procura	
Parâmetros	população inicial[a]	limite inferior	limite superior
$\ln k_1$(673 K)	−4,064	−10	1
$\ln k_2$(673 K)	−4,720	−10	1
$\ln k_3$(673 K)	−3,912	−10	1
$\ln k_4$(673 K)	−4,557	−10	1
$\ln k_1$(798 K)	0,287	−10	1
$\ln k_2$(798 K)	−0,120	−10	1
$\ln k_3$(798 K)	−0,297	−10	1
$\ln k_4$(798 K)	−0,998	−10	1
t_0(723 K)/min	4,406	0	10
b/min kJ mol^{-1}	−103,2	−1000	0

[a] Dados recolhidos em Bates e Watts (1988, pp. 194 e 195). Os valores das constantes cinéticas foram extraídos da tabela 5.8, enquanto os de t_0(723 K) e b foram extraídos da tabela 5.9.

Tabela 5.6 Pirólise do xisto betuminoso: estimativas para os parâmetros do modelo

	Estimadores					
Parâmetros	ML	LAD	M	MML	MTL25	MTL50
$\ln k_1$(673 K)	−3,898	−3,967	−3,905	−4,101	−3,953	−4,045
$\ln k_2$(673 K)	−4,026	−4,048	−4,095	−4,745	−4,072	−4,234
$\ln k_3$(673 K)	−3,982	−3,885	−3,926	−4,122	−3,906	−3,927
$\ln k_4$(673 K)	−4,928	−4,895	−4,85	−4,622	−4,946	−4,798
$\ln k_1$(798 K)	0,5411	0,7124	0,6473	0,8552	1	0,9366
$\ln k_2$(798 K)	−0,02497	−0,05048	−0,05737	0,5713	0,5981	0,8537
$\ln k_3$(798 K)	0,02786	0,02	0,06201	0,5133	0,4396	0,66
$\ln k_4$(798 K)	−0,7332	−0,4425	−0,4365	−0,3376	−0,6446	−0,7298
t_0(723 K)/min	4,282	4,483	4,552	4,843	4,505	4,47
b/min kJ mol^{-1}	−95,62	−105,8	−112	−122,1	−66,43	−111

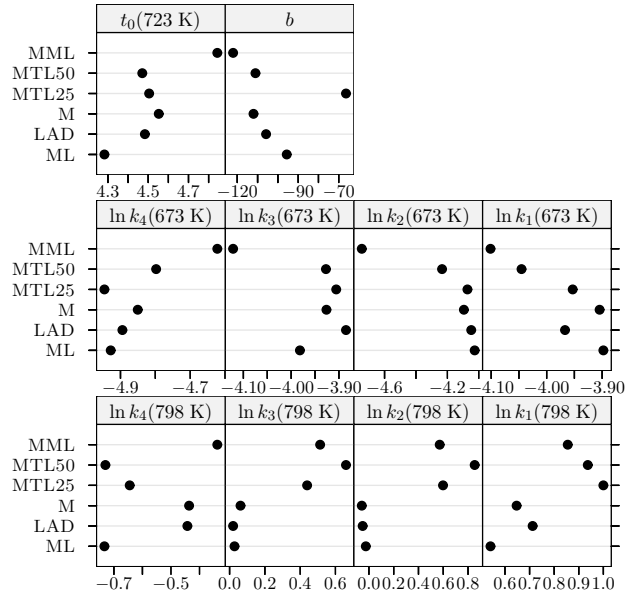


Figura 5.1 Pirólise do xisto betuminoso: estimativas para os parâmetros do modelo.

No que diz respeito aos erros de medição $\epsilon_k, k = (c_{\mathrm{B}}, c_{\mathrm{O}})$, a matriz das estimativas das covariâncias obtida para o estimador ML (critério do determinante) é $\widehat{\mathbf{\Sigma}} = \left[\begin{smallmatrix} 19{,}4 \\ -0{,}958 & 18{,}7 \end{smallmatrix}\right]$, sendo a correspondente correlação $-0{,}0504$.

5.3 Conversão do metanol em hidrocarbonetos

5.3.1 Descrição genérica do problema

Um mecanismo simplificado para a conversão do metanol em hidrocarbonetos encontra-se em Maria (1989),

$$\begin{aligned}
\mathrm{A} &\xrightarrow{k_1} \mathrm{B} \\
\mathrm{A} + \mathrm{B} &\xrightarrow{k_2} \mathrm{C} \\
\mathrm{C} + \mathrm{B} &\xrightarrow{k_3} \mathrm{P} \\
\mathrm{A} &\xrightarrow{k_4} \mathrm{C} \\
\mathrm{A} &\xrightarrow{k_5} \mathrm{P} \\
\mathrm{A} + \mathrm{B} &\xrightarrow{k_6} \mathrm{P}
\end{aligned}$$

onde A denota álcoóis e éteres, B metileno, C alcenos, e P alcanos, compostos aromáticos e outros produtos; k refere-se às constantes de velocidade dos diferentes passos. De notar que os passos 4, 5, e 6 são fictícios e, por isso, irrelevantes para esta reacção; a razão subjacente à sua inclusão foi testar os procedimentos de redução de modelos propostos

Tabela 5.7 Variação das fracções molares das espécies A, C, e P ao longo do tempo. Dados coligidos em Floudas *et al.* (1999, p. 402)

Obs.	t/s	x_A	x_C	x_P
1	0,050	0,461	0,114	0,018
2	0,065	0,426	0,135	0,035
3	0,080	0,383	0,157	0,045
4	0,123	0,305	0,194	0,047
5	0,233	0,195	0,231	0,084
6	0,273	0,170	0,234	0,095
7	0,354	0,139	0,228	0,111
8	0,397	0,112	0,228	0,134
9	0,418	0,112	0,226	0,168
10	0,502	0,090	0,220	0,148
11	0,553	0,082	0,214	0,157
12	0,681	0,066	0,178	0,206
13	0,750	0,053	0,188	0,206
14	0,916	0,043	0,183	0,214
15	0,937	0,041	0,184	0,213
16	1,122	0,029	0,166	0,230

nesse artigo.

Aplicando a hipótese do estado estacionário ao intermediário B, e considerando só reacções de primeira ordem (Maria, 1989), as equações diferenciais que determinam a variação com o tempo, t, das fracções molares das espécies, x, são dadas por

$$\frac{\mathrm{d}x_\mathrm{A}}{\mathrm{d}t} = -\Big(2k_1 - \frac{k_1 x_\mathrm{C}}{(k_{23}+k_{63})x_\mathrm{A} + x_\mathrm{C}} + k_4 + k_5\Big)x_\mathrm{A} \tag{5.5a}$$

$$\frac{\mathrm{d}c_\mathrm{C}}{\mathrm{d}t} = \frac{k_1 x_\mathrm{A}(k_{23}x_\mathrm{A} - x_\mathrm{C})}{(k_{23}+k_{63})x_\mathrm{A} + x_\mathrm{C}} + k_4 x_\mathrm{A} \tag{5.5b}$$

$$\frac{\mathrm{d}x_\mathrm{P}}{\mathrm{d}t} = \frac{k_1 x_\mathrm{A}(k_{63}x_\mathrm{A} + x_\mathrm{C})}{(k_{23}+k_{63})x_\mathrm{A} + x_\mathrm{C}} + k_5 x_\mathrm{A} \tag{5.5c}$$

com as condições iniciais $x_\mathrm{A} = 1$, $x_\mathrm{C} = 0$, e $x_\mathrm{P} = 0$ para $t = 0$, onde $k_{23} = k_2/k_3$ e $k_{63} = k_6/k_3$. Note-se que na formulação acima há apenas cinco parâmetros: k_1, k_{23}, k_4, k_5, e k_{63}.

A variação das fracções molares das espécies A, C e P com o tempo apresentam-se na tabela 5.7.

Na utilização da rotina de integração `LSODA`, estabeleceu-se 10^{-6} para as tolerâncias relativa (RTOL) e absoluta (ATOL). Na utilização do algoritmo MDE, estabeleceu-se $0,3$ como limite superior do espaço de procura dos elementos da matriz de covariâncias entre respostas para as simulações de Monte Carlo dos estimadores M, MML, e MTL.

Tabela 5.8 Conversão do metanol em hidrocarbonetos: parâmetros do algoritmo de optimização MDE usados na regressão dos dados experimentais

Espaço de procura $[0, 10^{-80}, 0, 0, 0] \leqslant [k_1, k_{23}, k_4, k_5, k_{63}] \leqslant [100, 100, 100, 100, 100]$					
N_P	F_b	C_R	ξ	Tolerância	Ponto incluído na população inicial
51	0,5	0,8	1,5	10^{-8}	$[5{,}2407, 1{,}2176, 0, 0, 0]$[a]

[a] Estimativas de mínimos quadrados apresentadas em Floudas *et al.* (1999, p. 402).

Tabela 5.9 Conversão do metanol em hidrocarbonetos: estimativas para os parâmetros do modelo

	Parâmetros				
Estimadores	k_1/s^{-1}	k_{23}/s^{-1}	k_4/s^{-1}	k_5/s^{-1}	k_{63}/s^{-1}
ML	3,1355	1,0084	$1{,}4252\times10^{-7}$	0,017516	0
LAD	4,0258	1,1172	$3{,}9269\times10^{-8}$	$1{,}665\times10^{-8}$	$1{,}8676\times10^{-10}$
M	4,4287	1,1552	$3{,}7839\times10^{-8}$	$7{,}9764\times10^{-10}$	$2{,}9875\times10^{-8}$
MML	91,233	1,4083	$1{,}6597\times10^{-9}$	$2{,}8932\times10^{-8}$	$2{,}4531\times10^{-10}$
MTL25	3,3092	1,025	$1{,}4328\times10^{-11}$	$1{,}5082\times10^{-13}$	$6{,}7818\times10^{-12}$
MTL50	2,9159	1,0249	$1{,}3616\times10^{-14}$	0,040191	$4{,}8216\times10^{-10}$

5.3.2 Aplicação aos dados experimentais

Os valores dos parâmetros de controlo do algoritmo MDE encontram-se na tabela 5.8. Na tabela 5.9 reproduzem-se as estimativas dos parâmetros do modelo (5.5) obtidas utilizando os dados experimentais.

Atendendo a que as três últimas etapas não contribuem para a reacção, é de antever que as estimativas relativas a k_4, k_5, e k_{63} sejam praticamente nulas. Com efeito, os resultados obtidos mostram que estas estimativas se situam à volta de 0 com excepção das correspondentes aos estimadores MTL50 e ML. Os resultados em relação a k_1 e k_{23} pouco diferem entre os vários estimadores à excepção da estimativa MML de k_1, a qual se encontra notoriamente afastada das restantes.

A matriz de covariâncias entre respostas $(x_\mathrm{A}, x_\mathrm{C}, x_\mathrm{P})$, $\widehat{\boldsymbol{\Sigma}}$, obtida para o estimador ML (critério do determinante) e a correspondente matriz de correlações, $\widehat{\boldsymbol{\Psi}}$, são

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0{,}015316 & & \\ 0{,}000177 & 0{,}0000332 & \\ -0{,}000793 & -0{,}0000419 & 0{,}000190 \end{bmatrix}, \quad \widehat{\boldsymbol{\Psi}} = \begin{bmatrix} 1 & & \\ 0{,}248 & 1 & \\ -0{,}464 & -0{,}527 & 1 \end{bmatrix}.$$

5.4 Hidrogenação catalítica do 3-hidroxipropanal

5.4.1 Descrição genérica do problema

Os dados da tabela 5.10 na próxima página dizem respeito ao processo de hidrogenação do 3-hidroxipropanal a 1,3-propanodiol catalisado por $Ni/Si_2/Al_2O_3$ (Zhu *et al.*, 1997). O 1,3-propanodiol é um monómero apelativo para a formação de polímeros de interesse industrial.

Em termos gerais o esquema cinético proposto é o seguinte:

$$\begin{aligned}
\mathrm{HPA} + \mathrm{H_2} &\xrightarrow{k_1} \mathrm{PD} \\
\mathrm{HPA} + \mathrm{PD} &\xrightarrow{k_2} \mathrm{acetal} + \mathrm{H_2O} \\
\mathrm{HPA} &\underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} \mathrm{Ac} + \mathrm{H_2O} \\
\mathrm{HPA} + \mathrm{Ac} &\xrightarrow{k_4} \text{4-oxo-1,7-heptanodiol}
\end{aligned}$$

em que HPA denota 3-hidroxipropanal, PD 1,3-propanodiol, e Ac acroleína.

O sistema de equações diferenciais que descreve a dependência temporal da concentração de 3-hidroxipropanal, 1,3-propanodiol, e acroleína, é

$$\frac{\mathrm{d}c_{\mathrm{HPA}}}{\mathrm{d}t} = -(r_1 + r_2)c_{\mathrm{cat}} - (r_3 + r_4 - r_{-3}) \tag{5.6a}$$

$$\frac{\mathrm{d}c_{\mathrm{PD}}}{\mathrm{d}t} = (r_1 - r_2)c_{\mathrm{cat}} \tag{5.6b}$$

$$\frac{\mathrm{d}c_{\mathrm{Ac}}}{\mathrm{d}t} = r_3 - r_4 - r_{-3}, \tag{5.6c}$$

com as condições iniciais $c_{\mathrm{HPA}} = 1{,}34953$ mol/l, $c_{\mathrm{PD}} = 0$, e $c_{\mathrm{Ac}} = 0$ para $t = 0$, onde c representa as concentrações (mol/l) das várias espécies, r denota as velocidades de reacção dos diferentes passos, t é o tempo, e c_{cat} representa a concentração de catalisador (10 g/l).

O conjunto das velocidades de reacção é o seguinte

$$r_1 = \frac{k_1 p c_{\mathrm{HPA}}}{\left(1 + \left(\frac{K_1 p}{\mathcal{H}}\right)^{1/2} + K_2 c_{\mathrm{HPA}}\right)^3 \mathcal{H}} \tag{5.7a}$$

$$r_2 = \frac{k_2 c_{\mathrm{PD}} c_{\mathrm{HPA}}}{1 + \left(\frac{K_1 p}{\mathcal{H}}\right)^{1/2} + K_2 c_{\mathrm{HPA}}} \tag{5.7b}$$

$$r_3 = k_3 c_{\mathrm{HPA}} \tag{5.7c}$$

$$r_{-3} = k_{-3} c_{\mathrm{Ac}} \tag{5.7d}$$

$$r_4 = k_4 c_{\mathrm{Ac}} c_{\mathrm{HPA}}, \tag{5.7e}$$

onde k refere-se às constantes de velocidade ($\mathrm{l^2/mol\ min\ g}$ para k_1 e k_2 ou $\mathrm{min^{-1}}$) dos diferentes passos, p é a pressão de hidrogénio (MPa) no reactor, K_1 e K_2 são as constantes de adsorção (l/mol), respectivamente, do hidrogénio e do 3-hidroxipropanal,

Tabela 5.10 Variação com o tempo das concentrações de 3-hidroxipropanal e 1,3-propanodiol a 45 °C e diferentes pressões

$p = 2{,}6$ MPa				$p = 4{,}0$ MPa				$p = 5{,}15$ MPa			
Obs.	t/min	c_{HPA}/mol l^{-1}	c_{PD}/mol l^{-1}	Obs.	t/min	c_{HPA}/mol l^{-1}	c_{PD}/mol l^{-1}	Obs.	t/min	c_{HPA}/mol l^{-1}	c_{PD}/mol l^{-1}
1	10	1,37395	0,0	14	10	1,3295	0,00262812	27	10	1,36342	0,00262812
2	20	1,25821	0,0197109	15	20	1,31157	0,0525624	28	20	1,25882	0,0700394
3	30	1,18707	0,0642576	16	30	1,22828	0,120736	29	30	1,17918	0,184363
4	40	1,13292	0,136399	17	40	1,087	0,241393	30	40	0,972102	0,354008
5	50	1,03556	0,238633	18	50	0,994539	0,384888	31	50	0,825203	0,469777
6	60	0,961339	0,304599	19	60	0,811825	0,4682	32	60	0,697109	0,607359
7	80	0,734436	0,492378	20	80	0,600962	0,773193	33	80	0,421451	0,852431
8	100	0,564551	0,732326	21	100	0,386302	0,990802	34	100	0,232296	1,03535
9	120	0,374385	0,887254	22	120	0,204222	1,14954	35	120	0,128095	1,16413
10	140	0,214799	1,04284	23	140	0,0782304	1,28	36	140	0,0289817	1,30053
11	160	0,100976	1,17306	24	160	0,0277708	1,29	37	160	0,00962368	1,31971
12	180	0,0364192	1,25769	25	180	0,00316296	1,30				
13	200	0,00530892	1,26032	26	200	0,00210864	1,30				

Fonte: Tabela 16.23 de Englezos e Kalogerakis (2001, p. 309).

Tabela 5.11 Hidrogenação catalítica do 3-hidroxipropanal: dados para o algoritmo MDE

	Ponto incluído na	Espaço de procura	
Parâmetros	população inicial[a]	limite inferior	limite superior
k_1	6,533	0	100
k_2	$3{,}048\times10^{-4}$	0	0,1
k_3	$6{,}233\times10^{-6}$	0	0,1
k_{-3}	$7{,}219\times10^{-4}$	0	0,1
k_4	$3{,}902\times10^{-6}$	0	0,1
K_1	95,00	0	1000
K_2	3,227	0	100

[a] Tabela 3 de Zhu *et al.* (1997, p. 2900).

Tabela 5.12 Hidrogenação catalítica do 3-hidroxipropanal: estimativas para os parâmetros do modelo

	Parâmetros						
Estimadores	k_1	k_2	k_3	k_{-3}	k_4	K_1	K_2
ML	12,07	$4{,}031\times10^{-10}$	0,0006636	0,0169	0,015	169,2	4,219
LAD	7,053	$1{,}991\times10^{-10}$	0,0005595	0,008546	0,01461	114,7	3,358
M	11,58	$7{,}076\times10^{-10}$	0,0003891	$7{,}184\times10^{-8}$	0,02786	167,1	4,121
MML	8,676	$1{,}813\times10^{-5}$	0,0003362	0,001028	0,001635	184,5	3,208
MTL25	4,627	$5{,}036\times10^{-13}$	0,0003946	0,002376	0,02277	88,43	2,717
MTL50	21,26	$5{,}453\times10^{-12}$	0,0009651	0,01359	0,01071	168,8	5,754

$[k_1, k_2] = \mathrm{l}^2/\mathrm{mol\ min\ g}$; $[k_3, k_{-3}, k_4] = \mathrm{min}^{-1}$; $[K_1, K_2] = \mathrm{l/mol}$

e $\mathcal{H}$ é constante da lei de Henry, que a 25 °C toma o valor de $\mathcal{H} = 137{,}9$ MPa l/mol. Os parâmetros do modelo organizam-se no vector $\boldsymbol{\theta} = [k_1, k_2, k_3, k_{-3}, k_4, K_1, K_2]^T$.

Na utilização da rotina de integração `LSODA`, estabeleceu-se 10^{-6} para as tolerâncias relativa (RTOL) e absoluta (ATOL). Na utilização do algoritmo MDE, estabeleceu-se 1 como limite superior do espaço de procura dos elementos da matriz de covariâncias entre respostas para as simulações de Monte Carlo dos estimadores M, MML, e MTL.

5.4.2 Aplicação aos dados experimentais

Os valores usados para os parâmetros de controlo do algoritmo MDE são: $N_\mathrm{P} = 51$, $F_\mathrm{b} = 0{,}5$, $C_\mathrm{R} = 0{,}8$, e $\xi = 1{,}5$; a tolerância usada para o critério de paragem foi 10^{-8}. A tabela 5.11 indica o espaço de procura e as estimativas incluidas na população inicial.

A matriz de covariâncias entre respostas ($c_\mathrm{HPA}, c_\mathrm{PD}$), obtida para o estimador ML (critério do determinante) é $\widehat{\boldsymbol{\Sigma}} = \left[\begin{smallmatrix} 0{,}00305 & \\ -0{,}00229 & 0{,}00295 \end{smallmatrix}\right]$, sendo a correspondente correlação $-0{,}763$.

Na tabela 5.12 encontram-se as estimativas dos parâmetros do modelo (5.6) obtidas a partir dos dados experimentais. Na figura 5.2 na página ao lado apresentam-se gráficos para os resultados referentes à tabela 5.12.

Os resultados obtidos mostram, para todos os parâmetros, uma variação clara entre as diferentes estimativas.

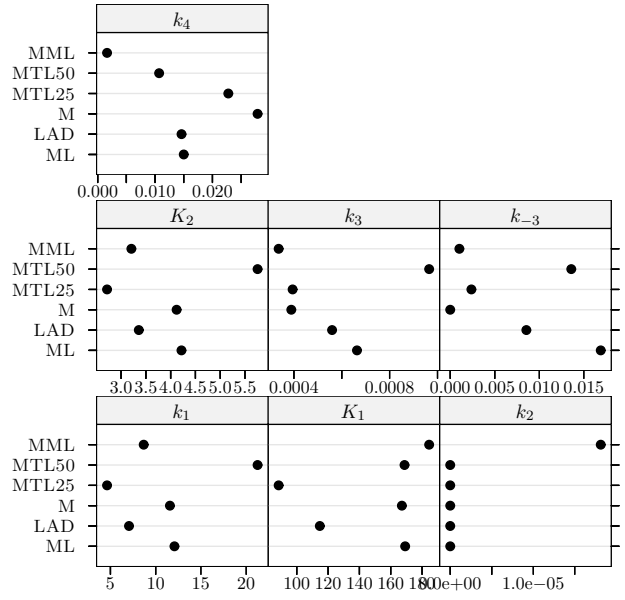


Figura 5.2 Hidrogenação catalítica do 3-hidroxipropanal: estimativas para os parâmetros do modelo.

Tabela 5.13 Tempos de cálculo dos diferentes problemas para o cenário sem *outliers*, obtidos com um processador Pentium 4 a 1,8 GHz e a versão R 1.4.0

Problema	Tempo de cálculo/h
Hidrogenação do tolueno	33,7
Pirólise do xisto betuminoso	79,6
Conversão do metanol em hidrocarbonetos	34,8
Hidrogenação catalítica do 3-hidroxipropanal	100,8

5.5 Resultados e discussão das experiências com dados simulados sem outliers

Recordemos os critérios usados para averiguar empiricamente o desempenho de amostra finita de um estimador, aliás já descritos na secção 3.3 na página 44. A saber:

1. medida de eficiência
$$\frac{\text{med}(|\hat{\theta}_{\text{ML}} - \theta|)}{\text{med}(|\hat{\theta} - \theta|)};$$
2. medida de enviesamento
$$\text{RB} = \text{med}(\hat{\theta}) - \theta.$$

Na mesma linha do caso univariado passamos agora a apresentar na tabela 5.13, para cada um dos problemas estudados, os tempos de cálculo requeridos no cenário actual.

Nota O comportamento estranho das medidas de desempenho no caso dos parâmetros k_4, k_5, e k_{63} da conversão do metanol em hidrocarbonetos e k_2 da hidrogenação catalítica do 3-hidroxipropanal decorre dos verdadeiros valores desses parâmetros serem

essencialmente nulos. Assim, é previsível que qualquer estimador mostre valores bastante elevados (em termos relativos) de dispersão ou enviesamento, desde que as estimativas se localizem abaixo do limite que de um ponto de vista prático significa que as estimativas são 0.

Critério de eficiência A inspecção dos resultados apresentados nas figuras 5.3 a 5.7 nas páginas 127–131 mostra que há evidências da superioridade dos estimadores robustos sobre o estimador do critério do determinante para as distribuições CN(0,30, 5) e t-Student, embora bastante fracas em t-Student. No entanto, contrariamente ao caso univariado, a análise dos dados revela uma imagem assaz mista no que diz respeito ao comportamento dos estimadores robustos.

Assim, para cada um dos casos da hidrogenação do tolueno, por uma parte, os estimadores MML e MTL50 apresentam um desempenho pior do que o de todos os outros em competição; por outra parte, os estimadores LAD e M mostram um desempenho bastante satisfatório (comparável ao do estimador do critério do determinante). Verifica-se ainda que o desempenho do estimador MTL25 é o melhor (conjuntamente com o estimador LAD) em CN(0,30, 5), intermédio sob a distribuição Gaussiana e em CN(0,10, 2), e basicamente idêntico ao dos estimadores LAD e M sob a distribuição t-Student. Grosso modo, o mesmo se passa com a pirólise do xisto betuminoso.

Olhemos agora para o caso da conversão do metanol em hidrocarbonetos. Aqui a hierarquização dos estimadores robustos inverte-se drasticamente; o desempenho dos estimadores MML e MTL é claramente superior ao dos estimadores M e LAD.

No caso da hidrogenação catalítica do 3-hidroxipropanal é de destacar o estimador MML, o qual apresenta em geral o índice de desempenho mais elevado; a classificação dos restantes estimadores robustos é inconclusiva.

Critério de enviesamento Comparemos agora os valores do índice de enviesamento robustificado, apresentados nas figuras 5.8 a 5.12 nas páginas 132–136. Pode-se verificar que há alguns casos com enviesamento bastante elevado. É o que se passa com a hidrogenação do tolueno para as estimativas MML e MTL50 com excepção das correspondentes ao parâmetro $k_{\mathrm{H},1}$, com a pirólise do xisto betuminoso para os parâmetros $\ln k_2(798\ \mathrm{K})$ e $\ln k_3(798\ \mathrm{K})$, e na generalidade dos casos da hidrogenação catalítica do 3-hidroxipropanal. Assim, parece razoável concluir que o grande contributo para as baixas eficiências apresentadas pelos estimadores MML e MTL50 no caso da hidrogenação do tolueno provém do efeito do enviesamento.

5.6 Resultados das experiências com dados simulados com outliers

Esta secção apresenta os resultados relativos ao caso de contaminação moderada (figuras 5.13 a 5.22 nas páginas 137–146), caso de contaminação severa (figuras 5.23 a 5.32 nas páginas 147–156), caso de forte contaminação e perturbação (figuras 5.33 a 5.42 nas páginas 157–166), e caso limite (figuras 5.43 a 5.52 nas páginas 167–176), para cada um dos problemas estudados.

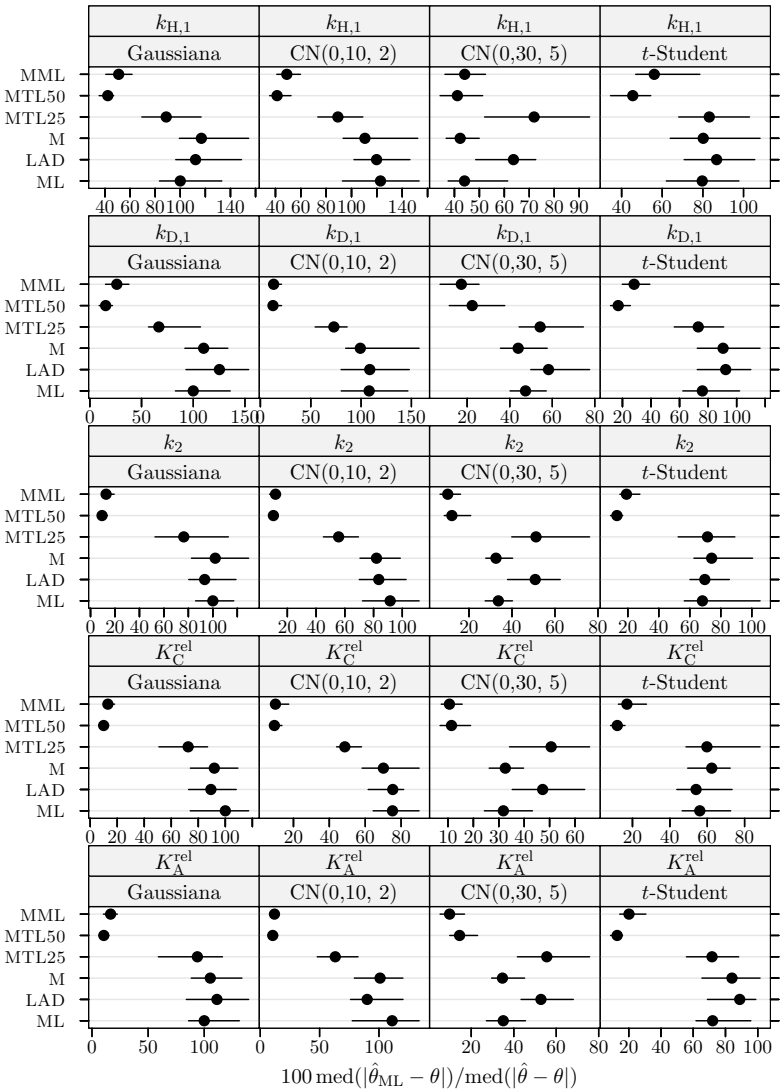


Figura 5.3 Hidrogenação do tolueno: medida da eficiência dos estimadores para dados simulados sem *outliers*. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

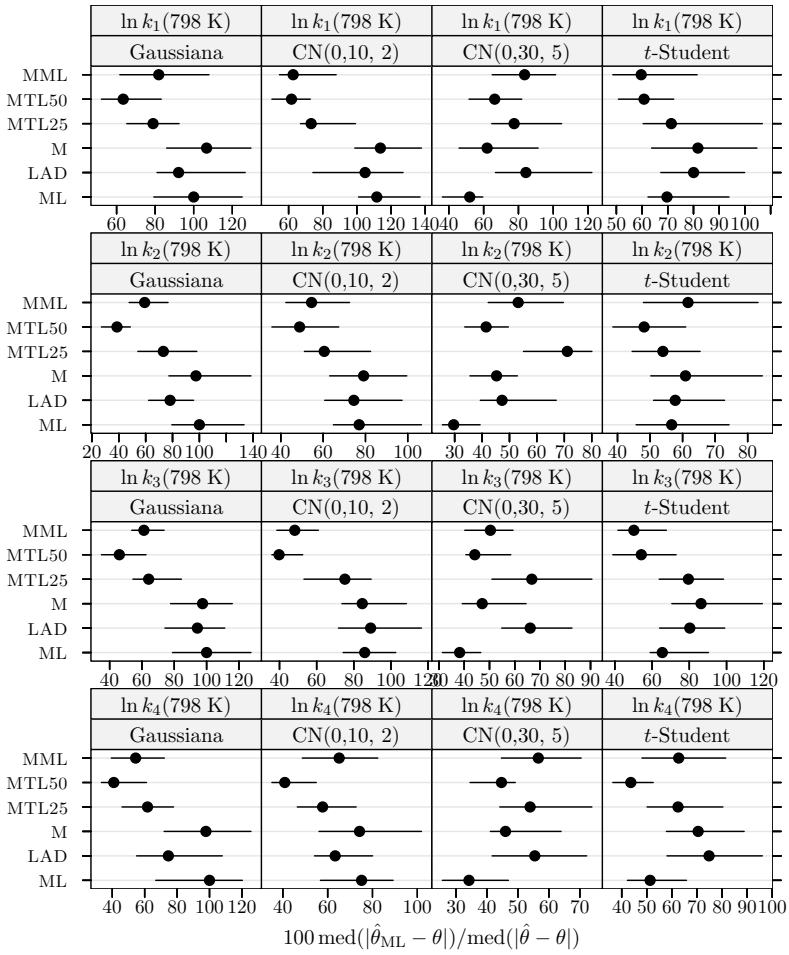


Figura 5.4 Pirólise do xisto betuminoso: medida da eficiência dos estimadores para dados simulados sem *outliers*. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

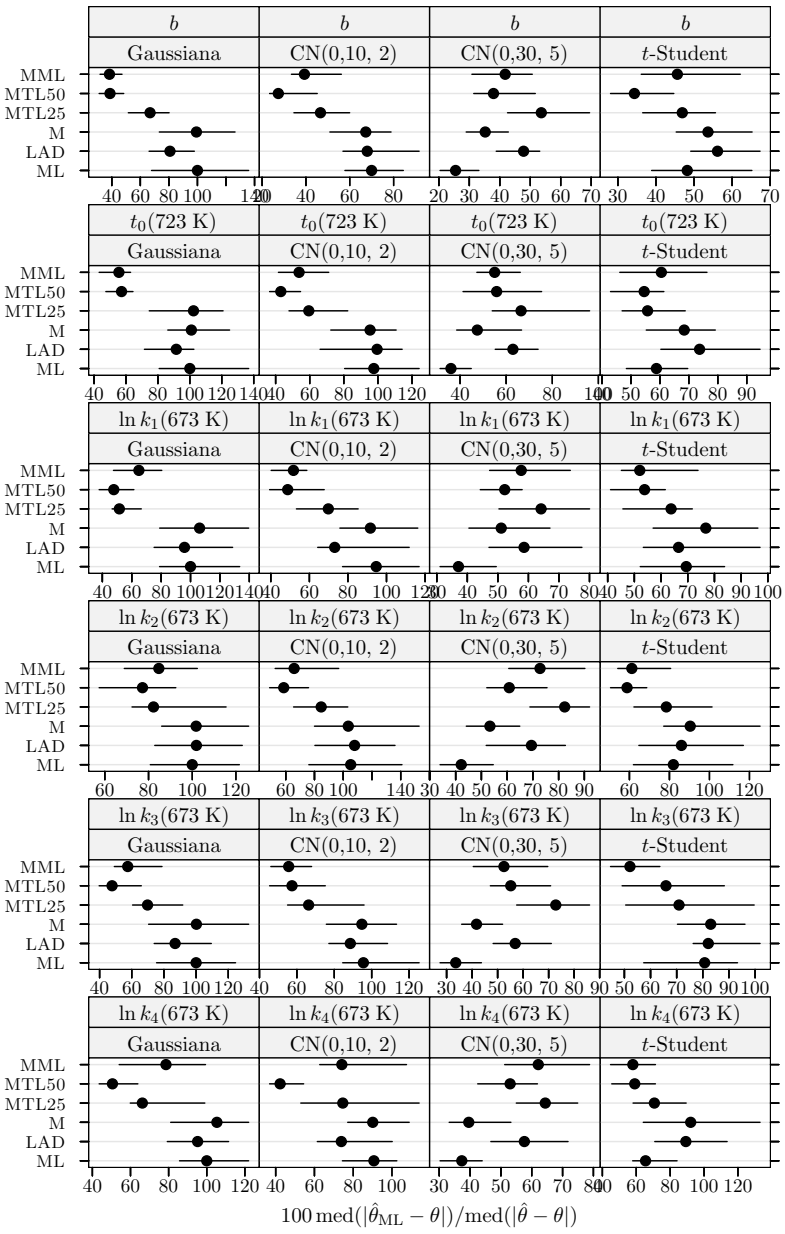


Figura 5.5 Pirólise do xisto betuminoso: medida da eficiência dos estimadores para dados simulados sem *outliers*. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

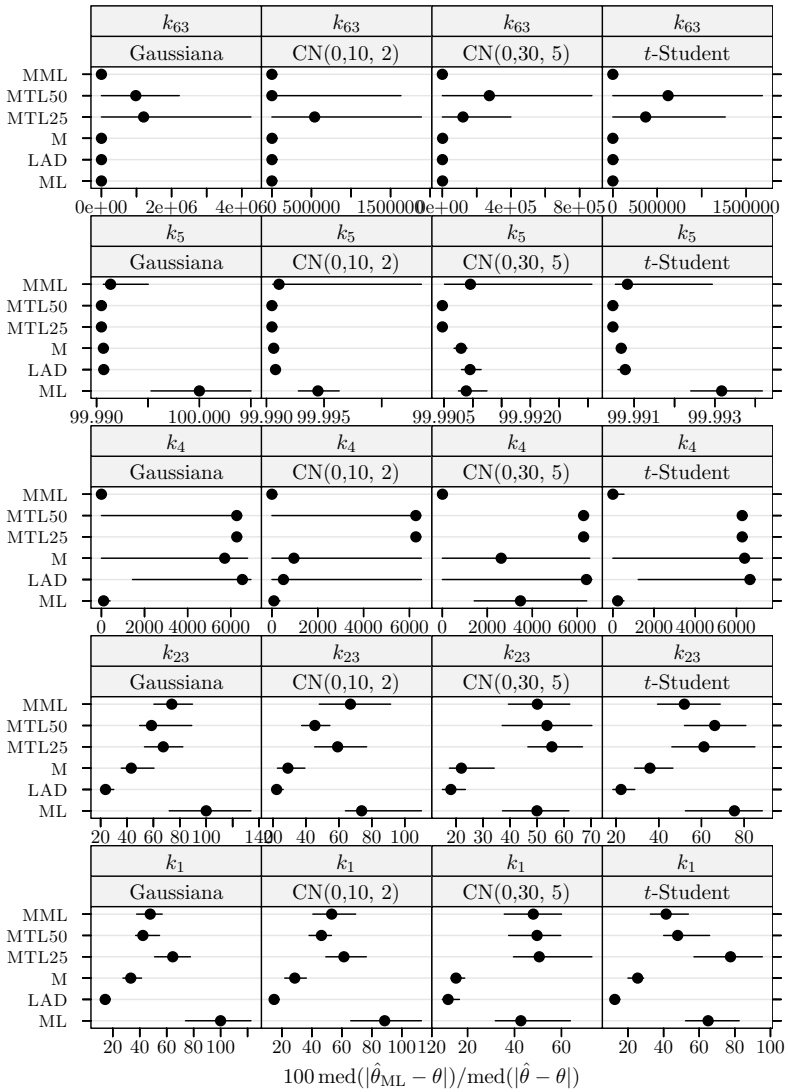


Figura 5.6 Conversão do metanol em hidrocarbonetos: medida da eficiência dos estimadores para dados simulados sem *outliers*. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

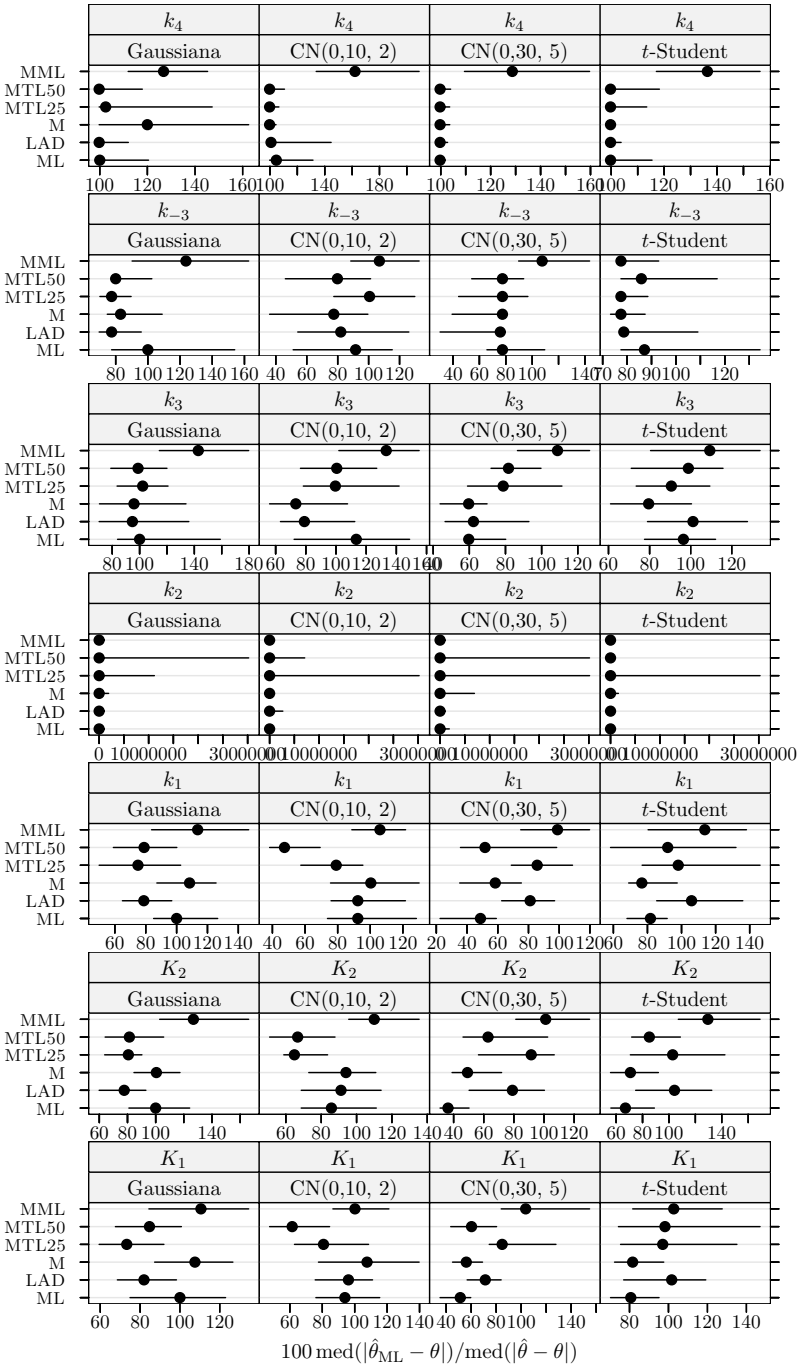


Figura 5.7 Hidrogenação catalítica do 3-hidroxipropanal: medida da eficiência dos estimadores para dados simulados sem *outliers*. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

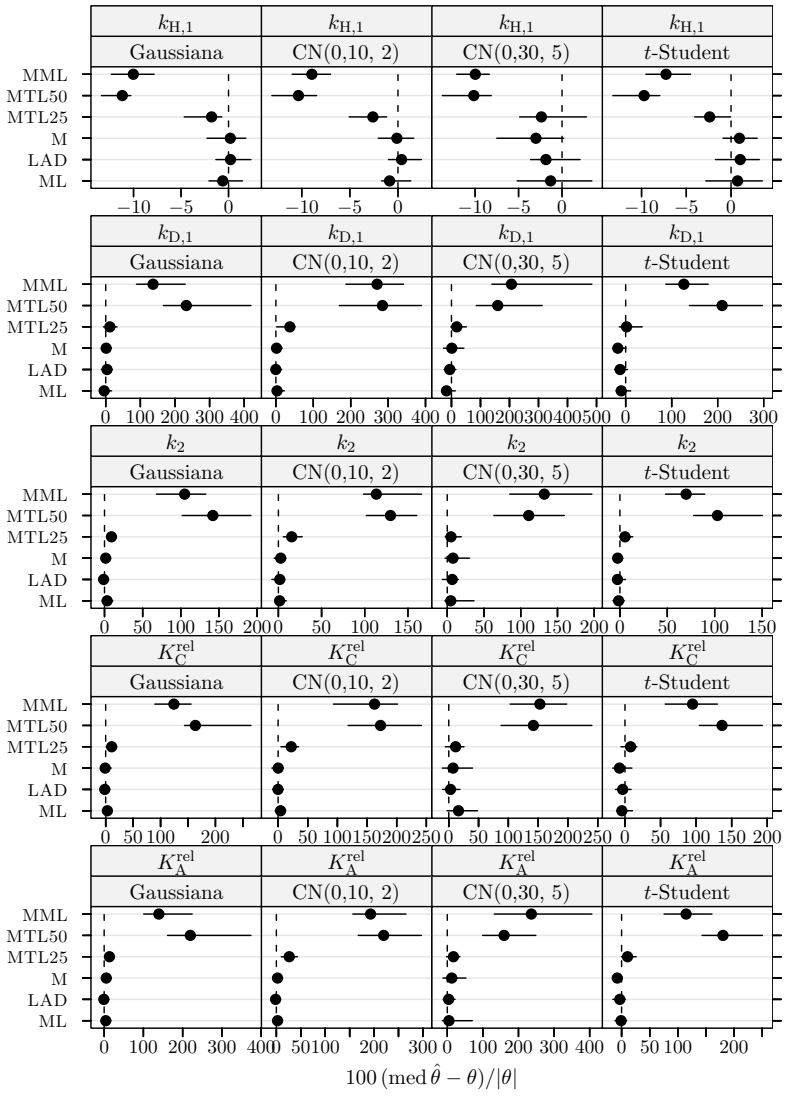


Figura 5.8 Hidrogenação do tolueno: índice de enviesamento robustificado dos estimadores para dados simulados sem *outliers*. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

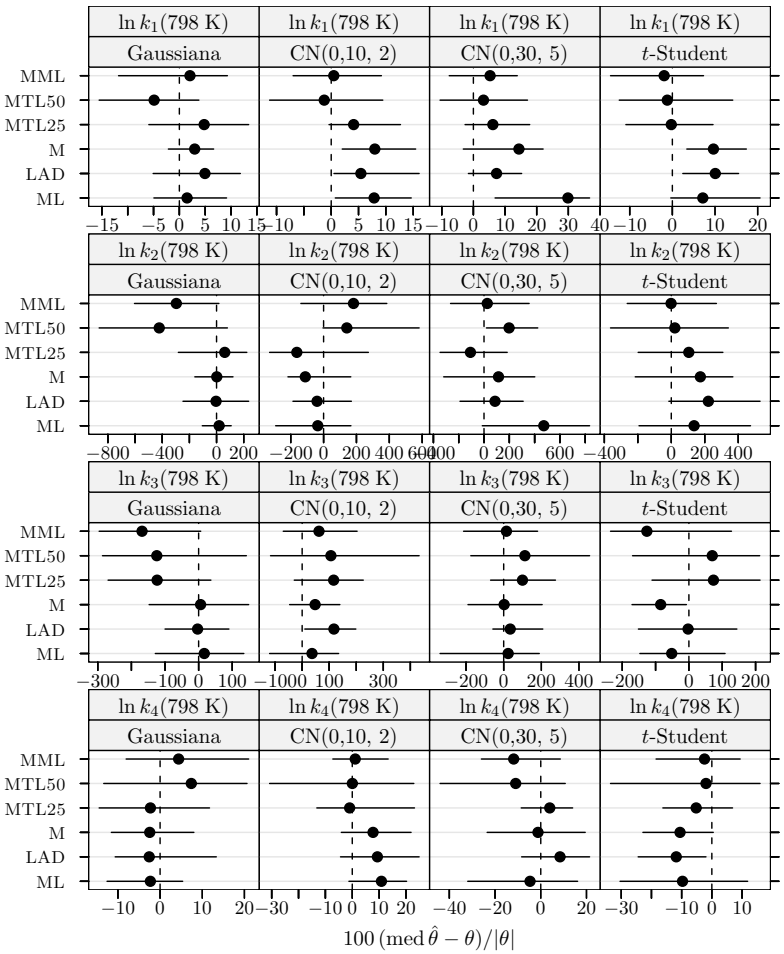


Figura 5.9 Pirólise do xisto betuminoso: índice de enviesamento robustificado dos estimadores para dados simulados sem *outliers*. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

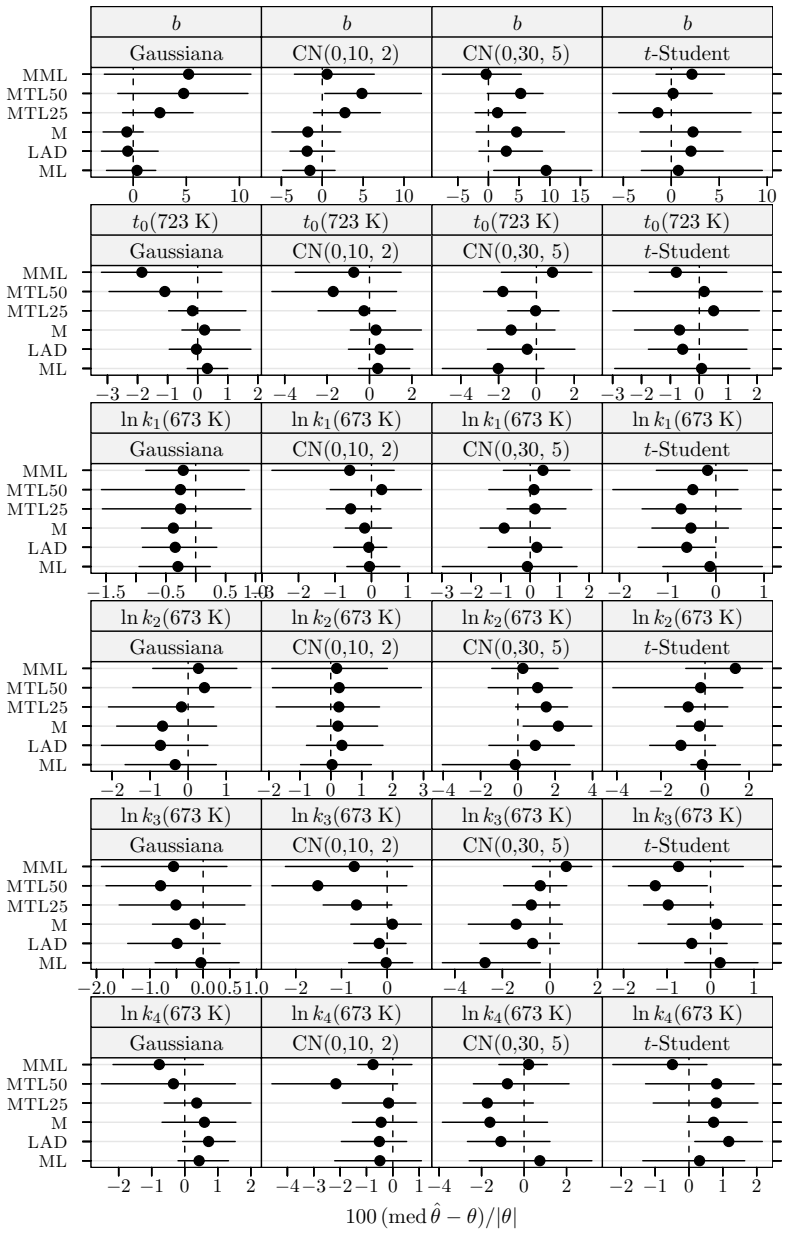


Figura 5.10 Pirólise do xisto betuminoso: índice de enviesamento robustificado dos estimadores para dados simulados sem *outliers*. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

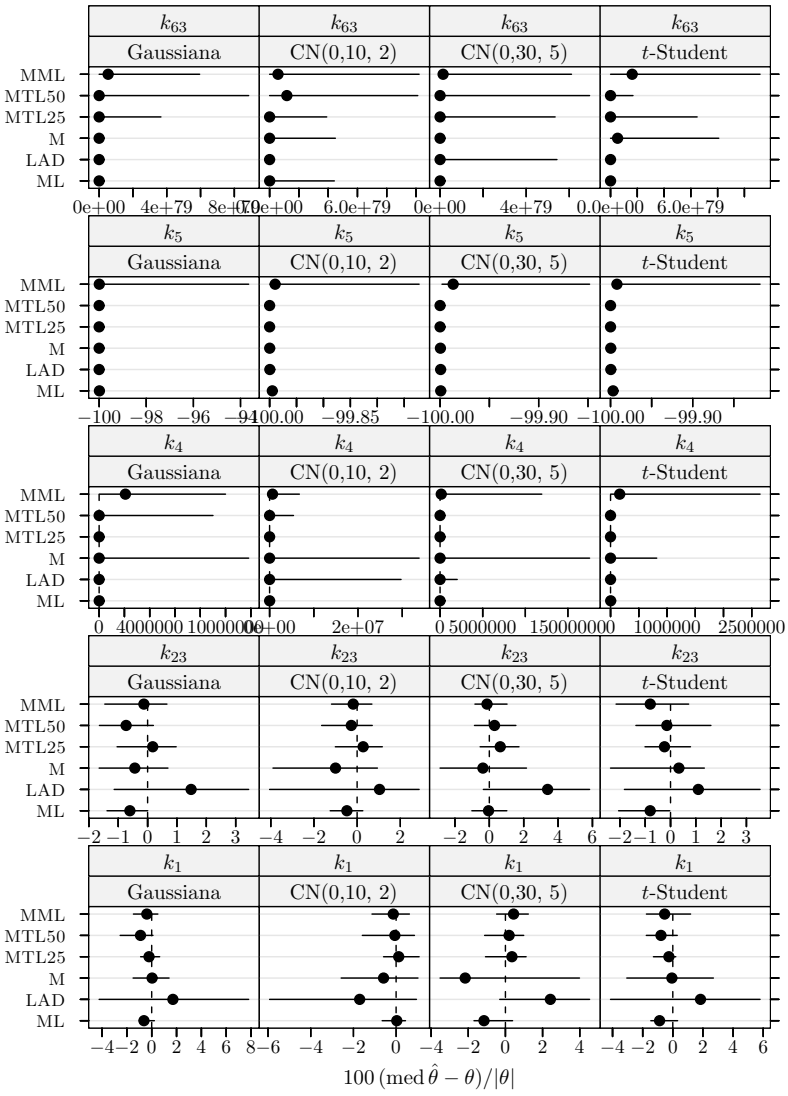


Figura 5.11 Conversão do metanol em hidrocarbonetos: índice de enviesamento robustificado dos estimadores para dados simulados sem *outliers*. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

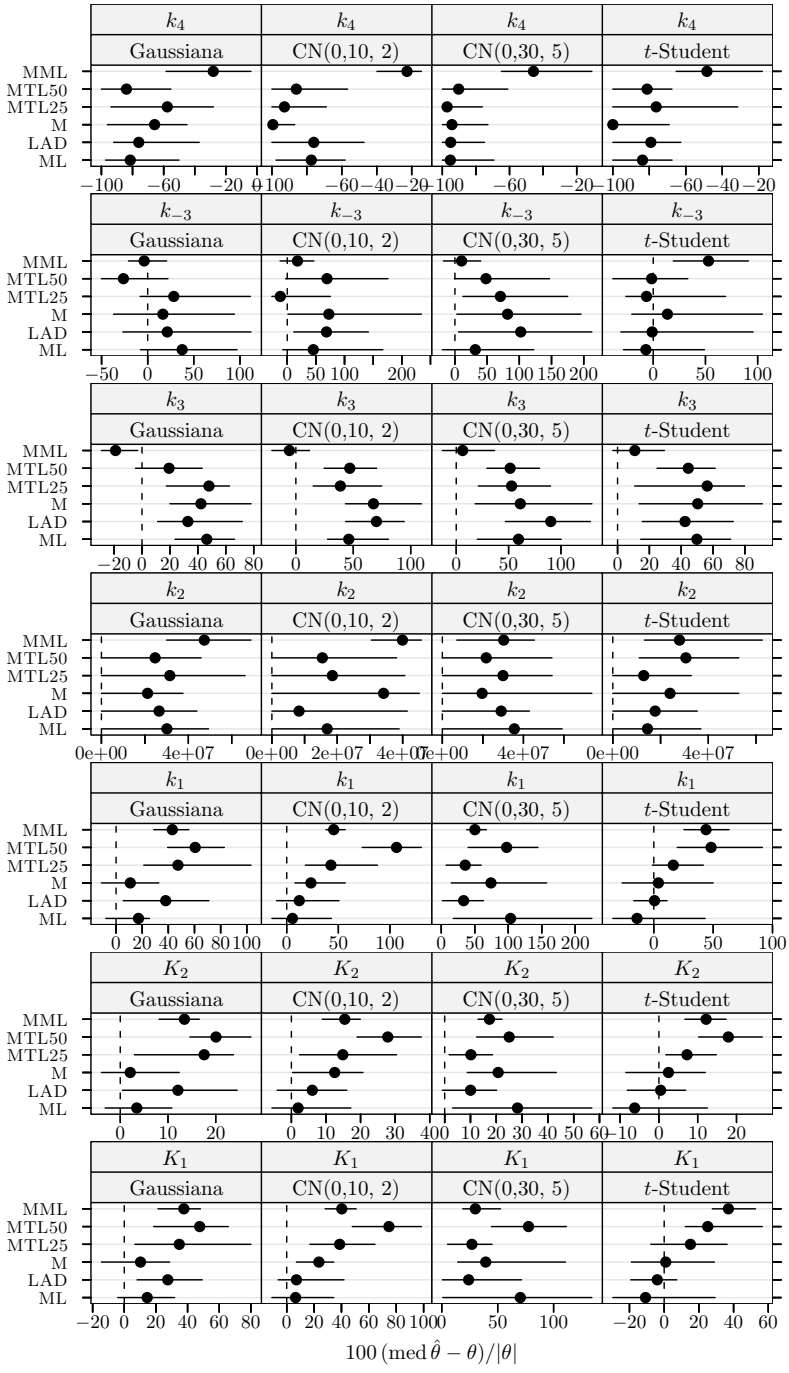


Figura 5.12 Hidrogenação catalítica do 3-hidroxipropanal: índice de enviesamento robustificado dos estimadores para dados simulados sem *outliers*. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras.

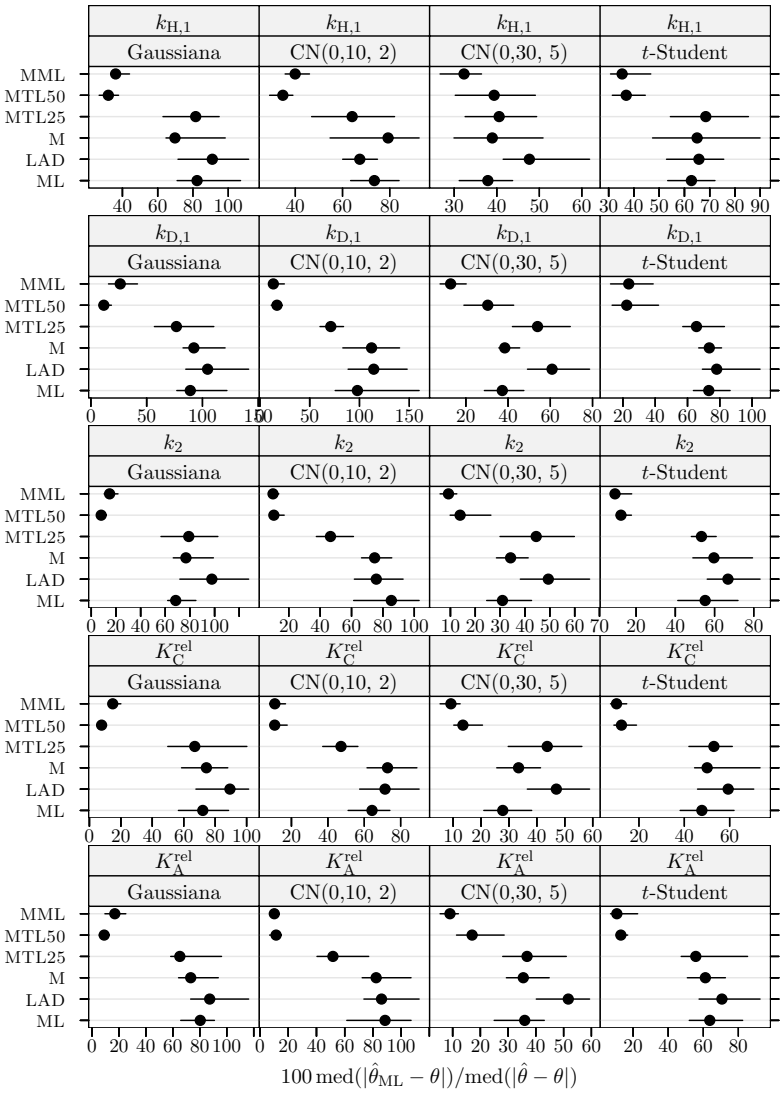


Figura 5.13 Hidrogenação do tolueno: medida da eficiência dos estimadores para dados simulados com 10% de *outliers* e $\delta_R = 5$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. O ponto perturbado foi 1.

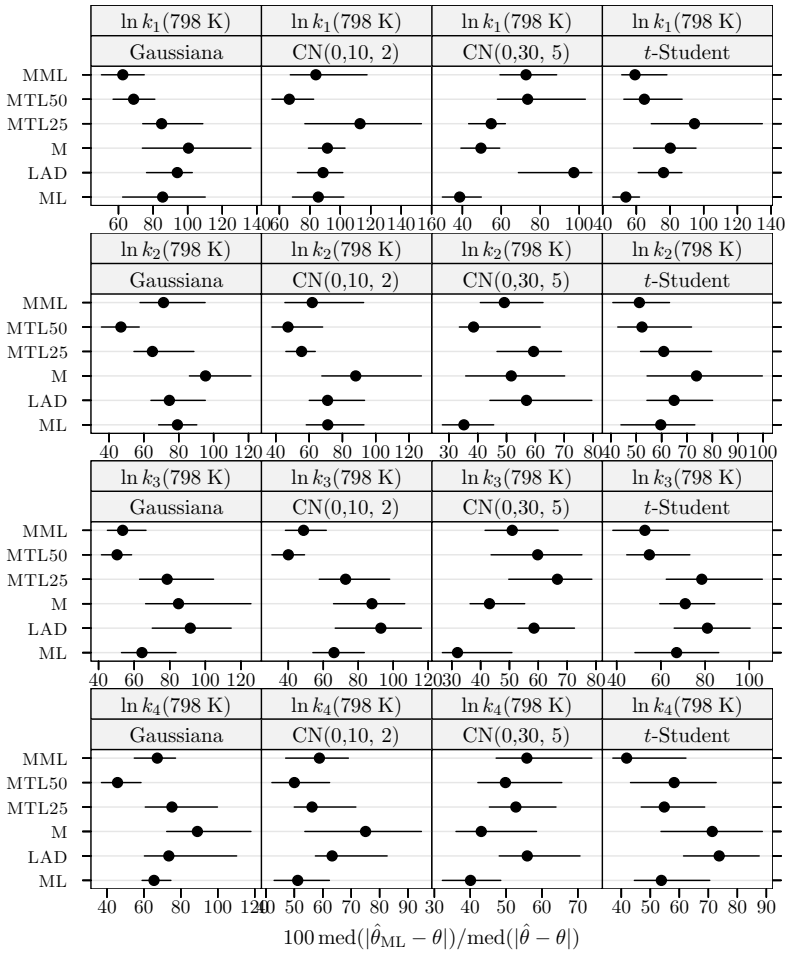


Figura 5.14 Pirólise do xisto betuminoso: medida da eficiência dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 4, 19, 32, 36, 37, e 42.

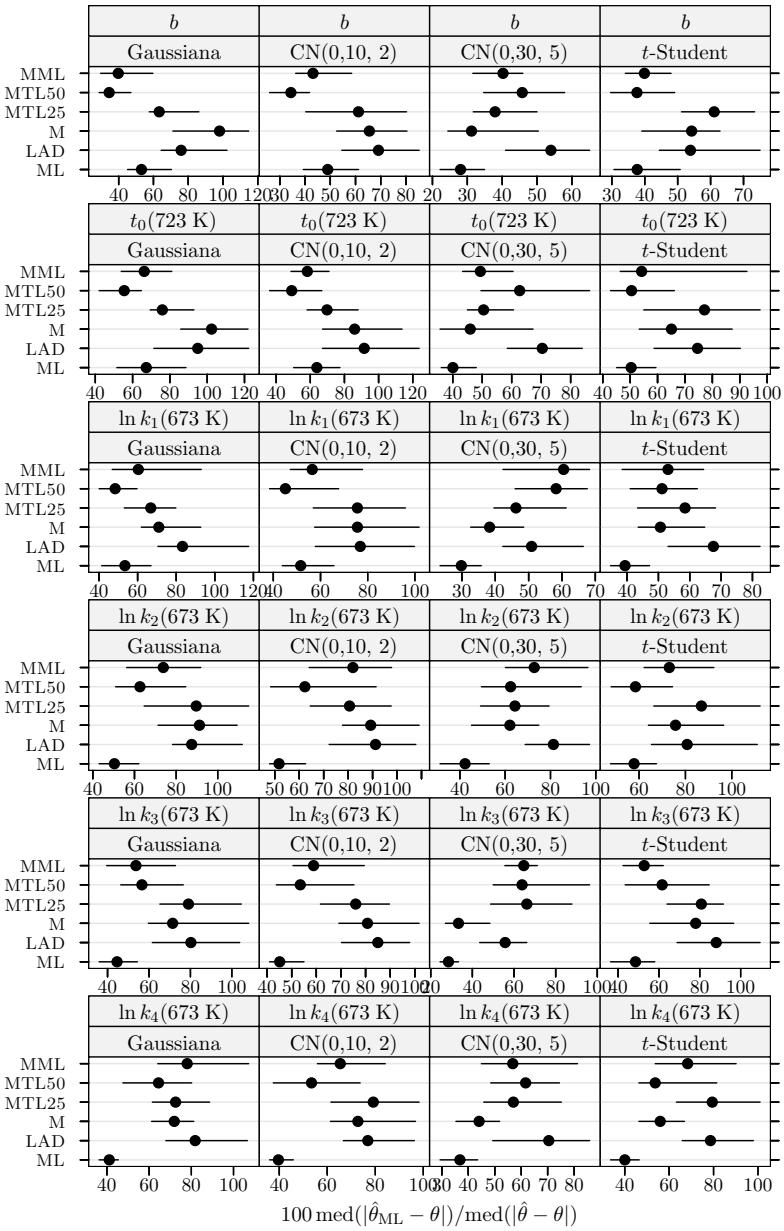


Figura 5.15 Pirólise do xisto betuminoso: medida da eficiência dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 4, 19, 32, 36, 37, e 42.

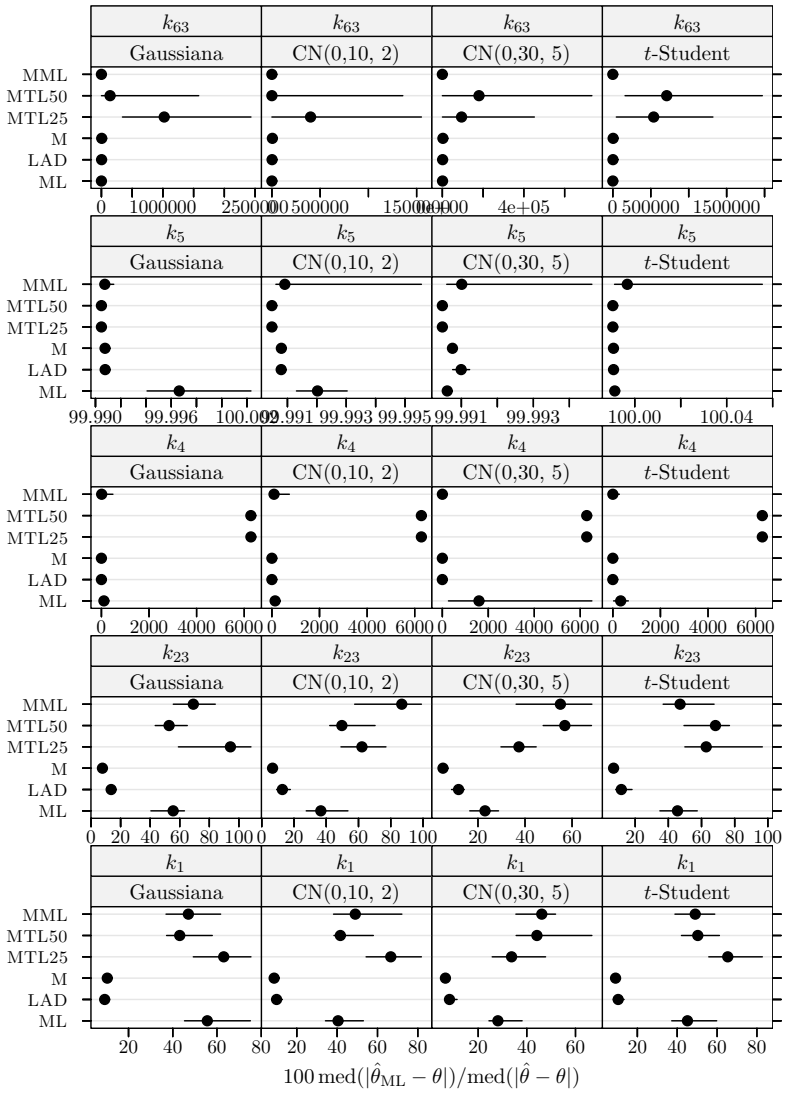


Figura 5.16 Conversão do metanol em hidrocarbonetos: medida da eficiência dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 5 e 11.

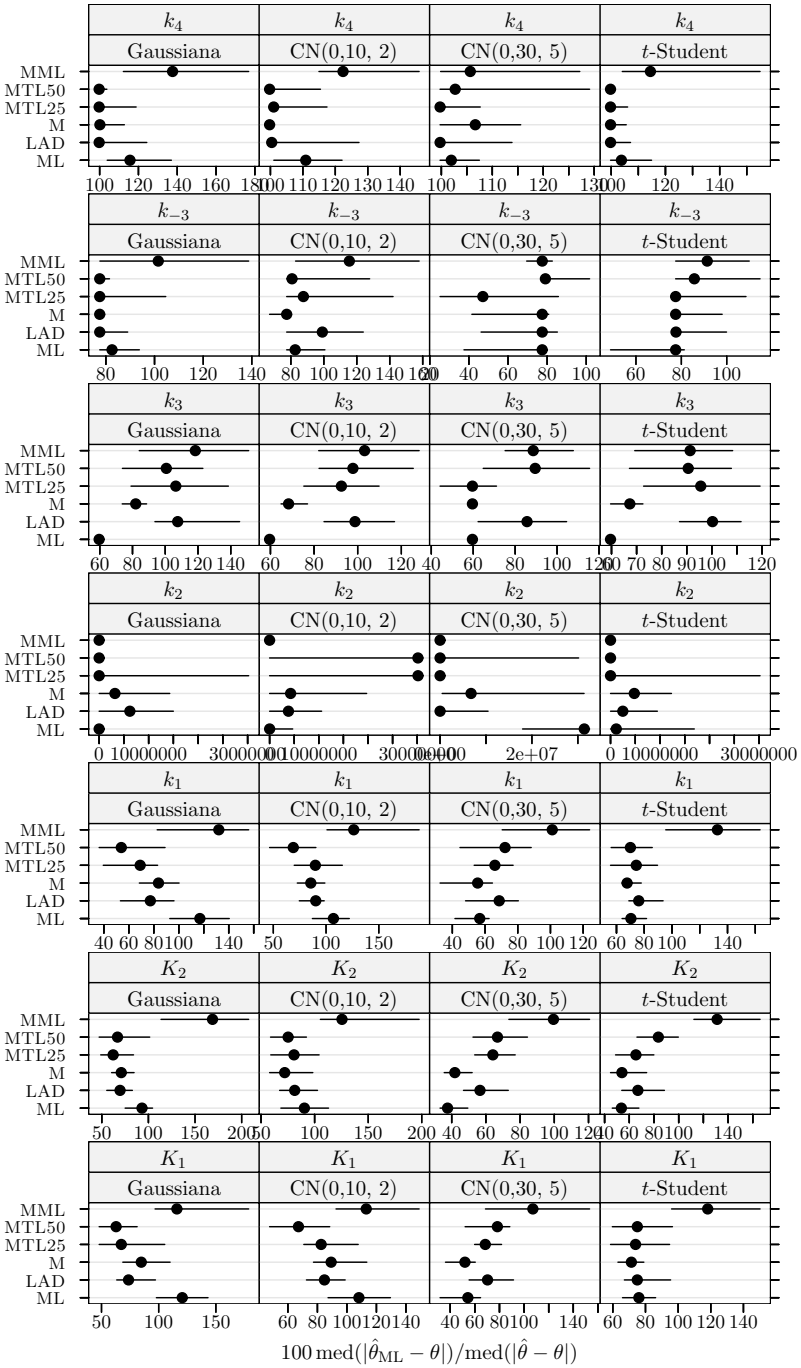


Figura 5.17 Hidrogenação catalítica do 3-hidroxipropanal: medida da eficiência dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 11, 23, 29, e 30.

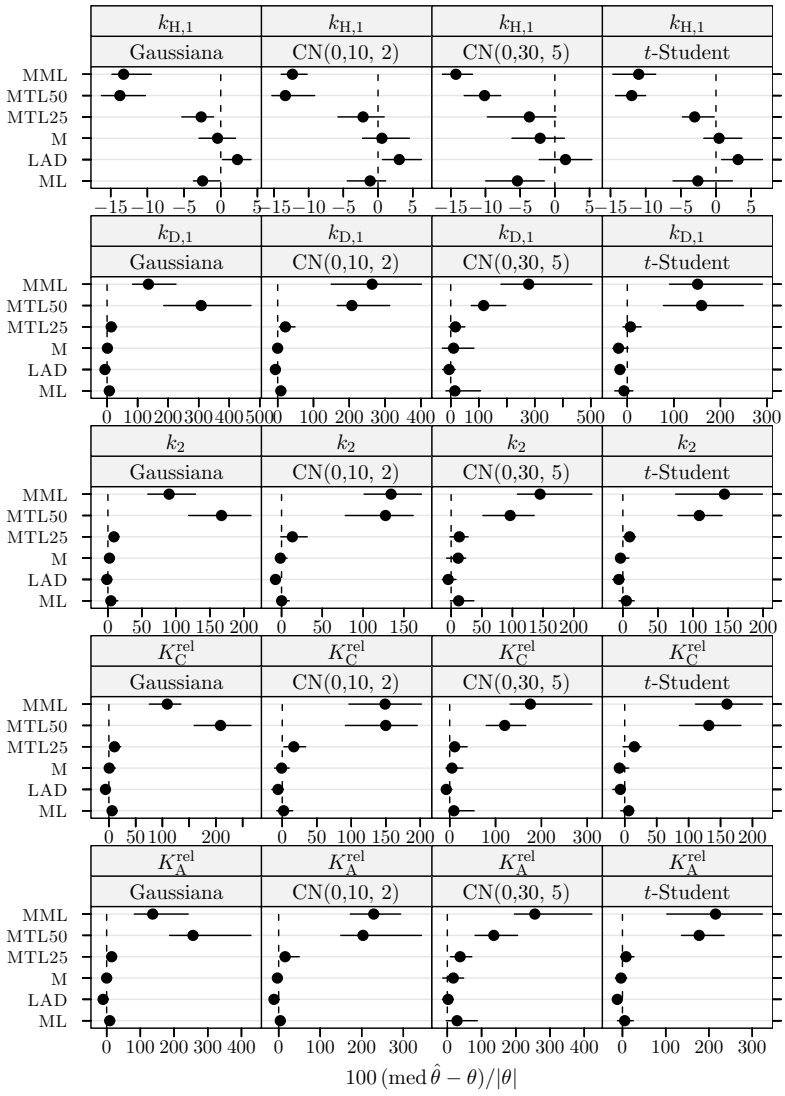


Figura 5.18 Hidrogenação do tolueno: índice de enviesamento robustificado dos estimadores para dados simulados com 10% de *outliers* e $\delta_R = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. O ponto perturbado foi 1.

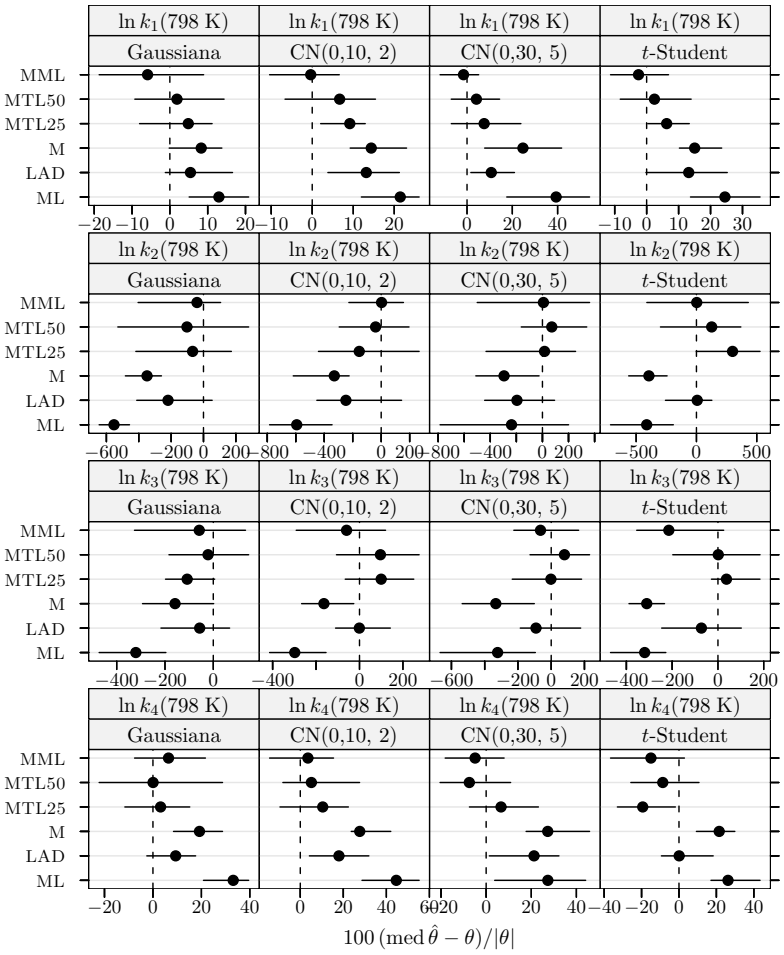


Figura 5.19 Pirólise do xisto betuminoso: índice de enviesamento robustificado dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 4, 19, 32, 36, 37, e 42.

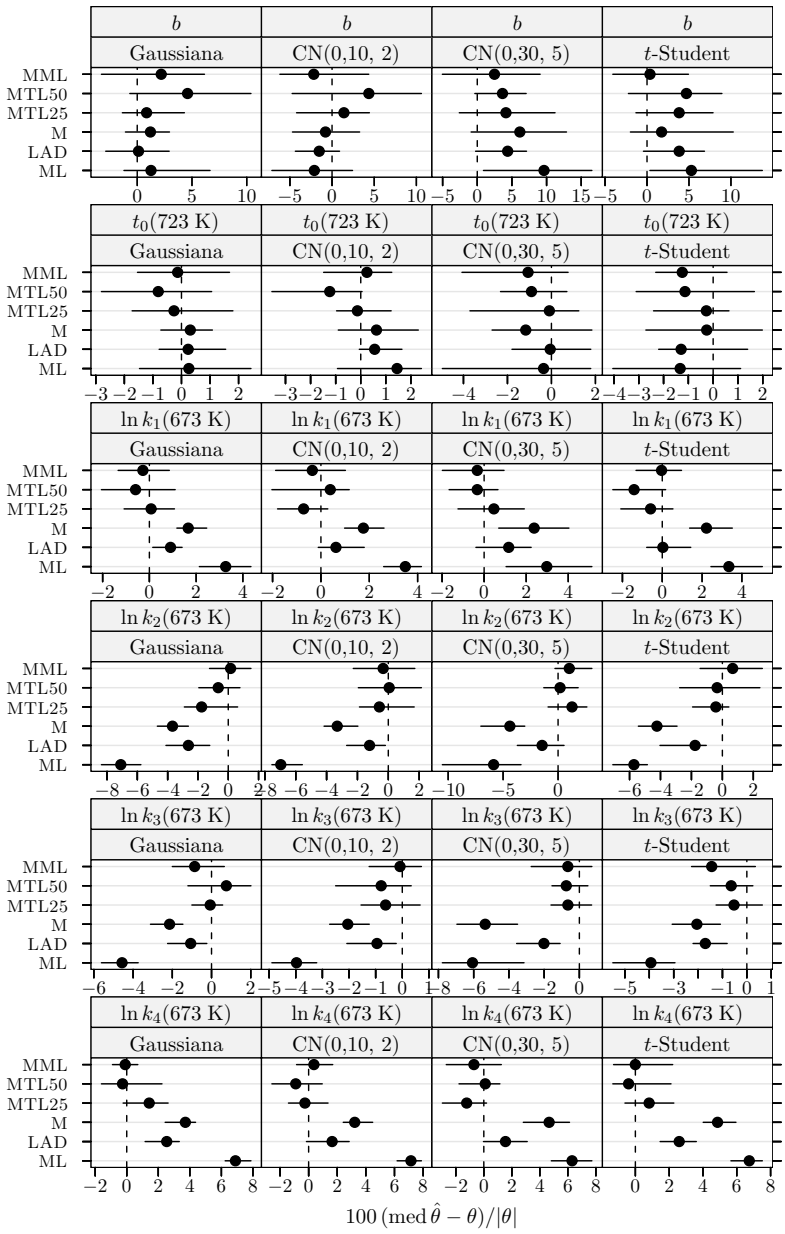


Figura 5.20 Pirólise do xisto betuminoso: índice de enviesamento robustificado dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 4, 19, 32, 36, 37, e 42.

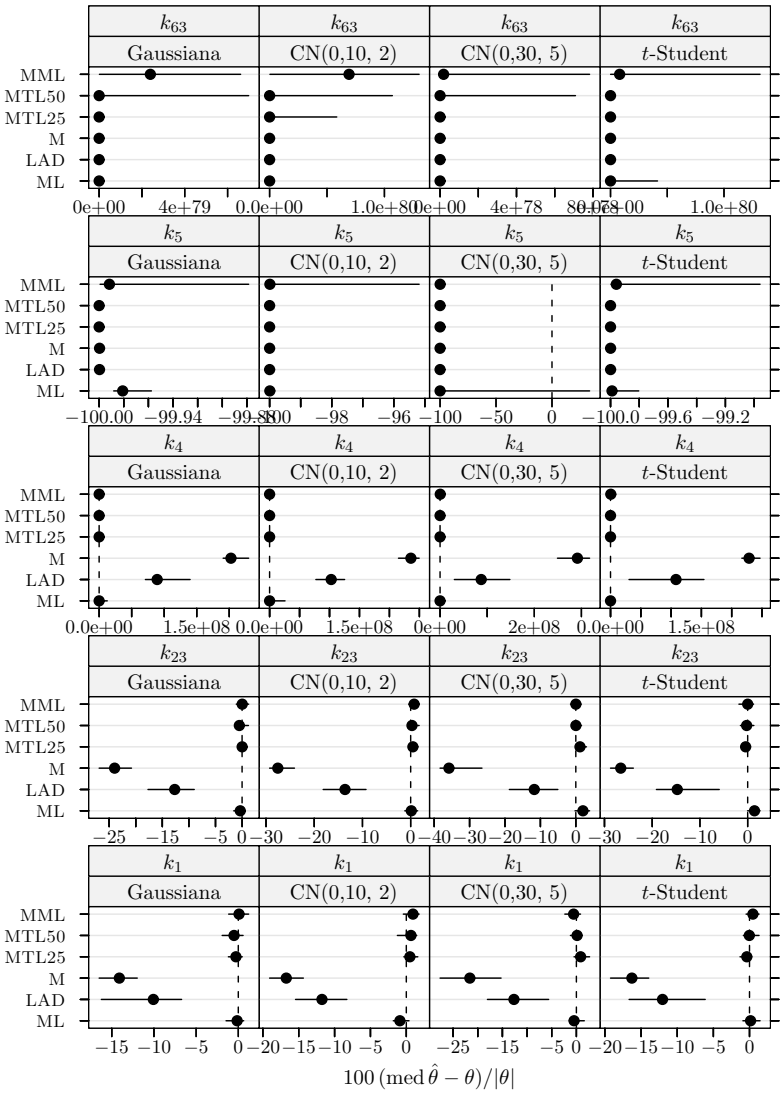


Figura 5.21 Conversão do metanol em hidrocarbonetos: índice de enviesamento robustificado dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 5 e 11.

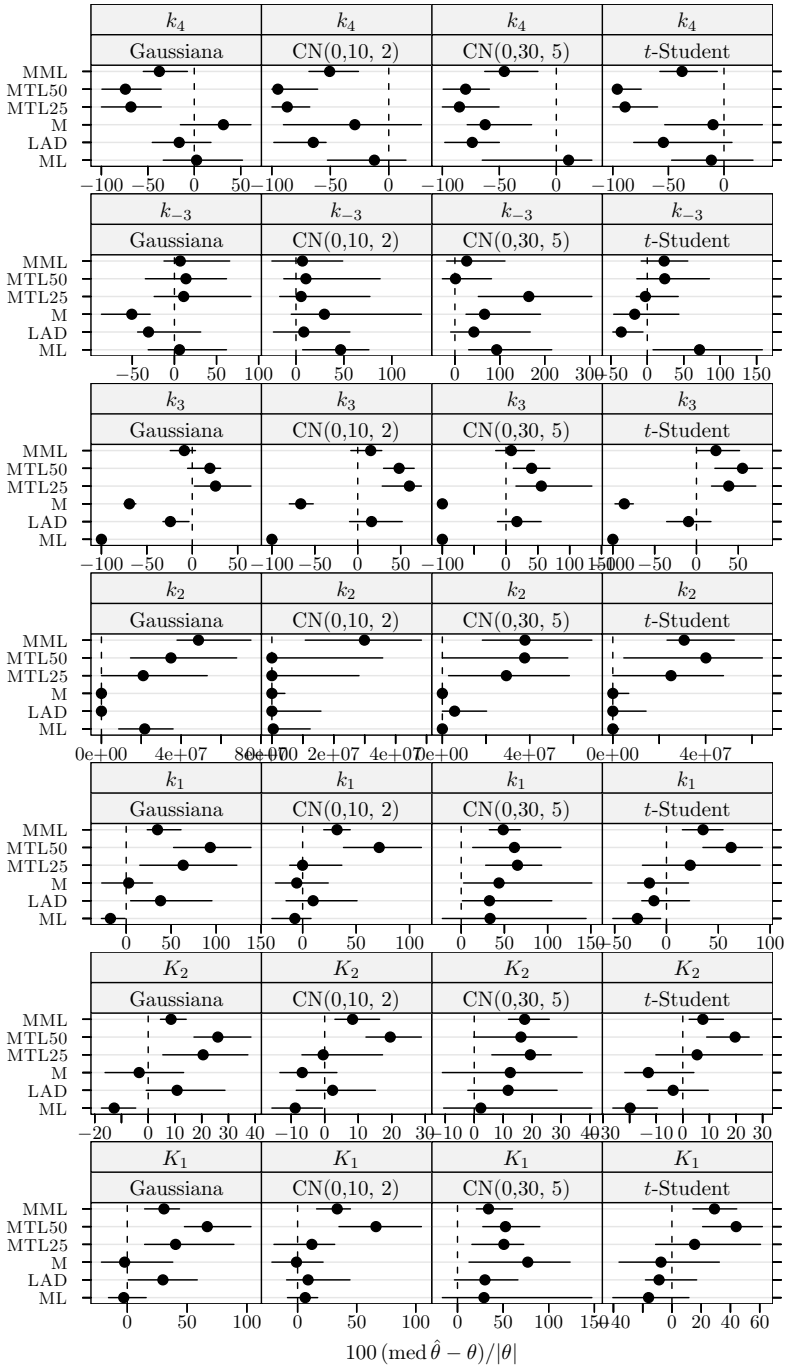


Figura 5.22 Hidrogenação catalítica do 3-hidroxipropanal: índice de enviesamento robustificado dos estimadores para dados simulados com 10% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 11, 23, 29, e 30.

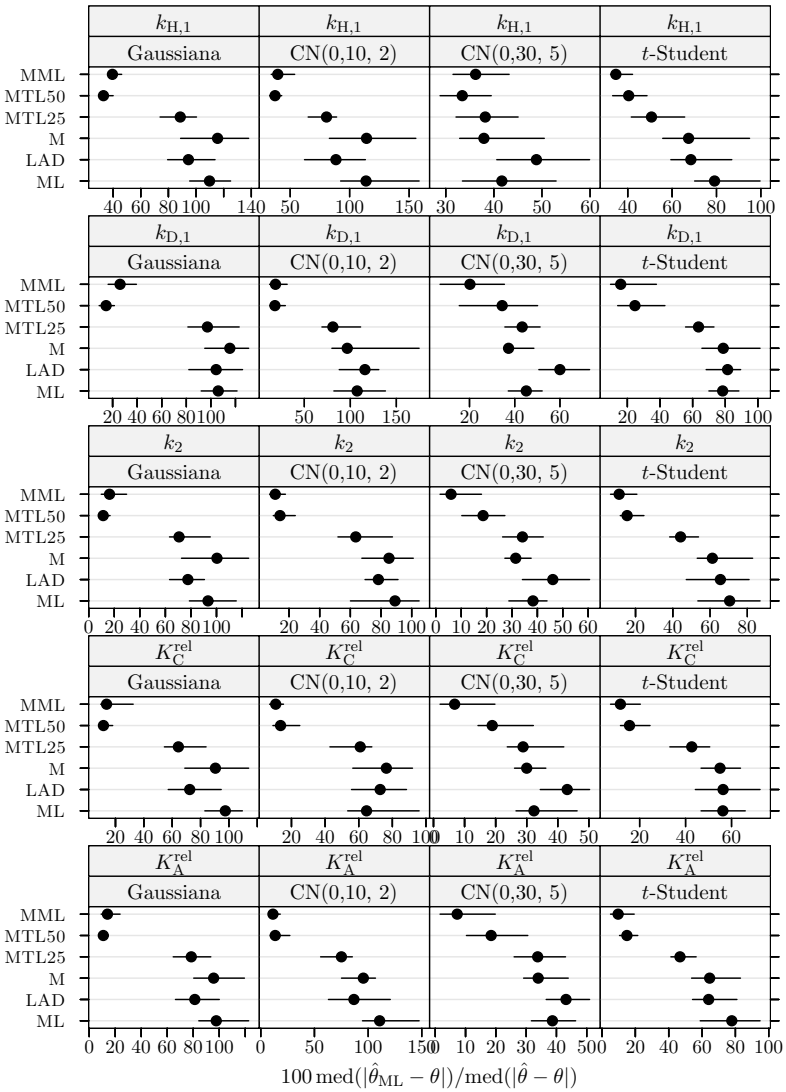


Figura 5.23 Hidrogenação do tolueno: medida da eficiência dos estimadores para dados simulados com 20% de *outliers* e $\delta_R = 5$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, e 6.

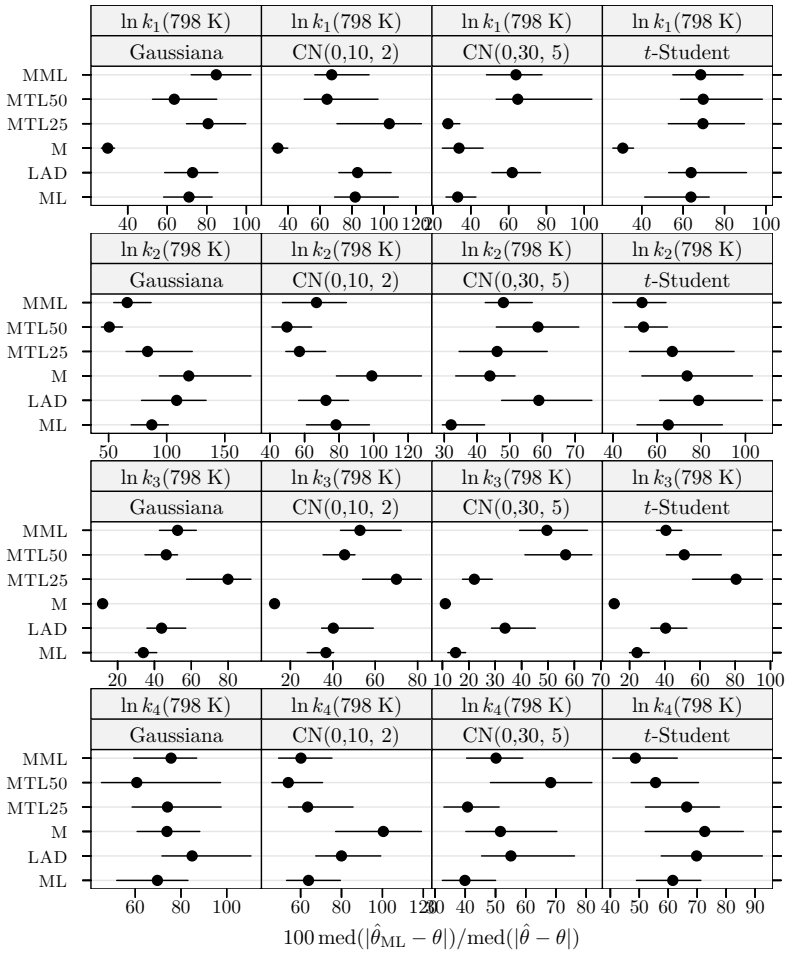


Figura 5.24 Pirólise do xisto betuminoso: medida da eficiência dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 10, 19, 20, 24, 32, 36, 37, 38, 42, 58, e 63.

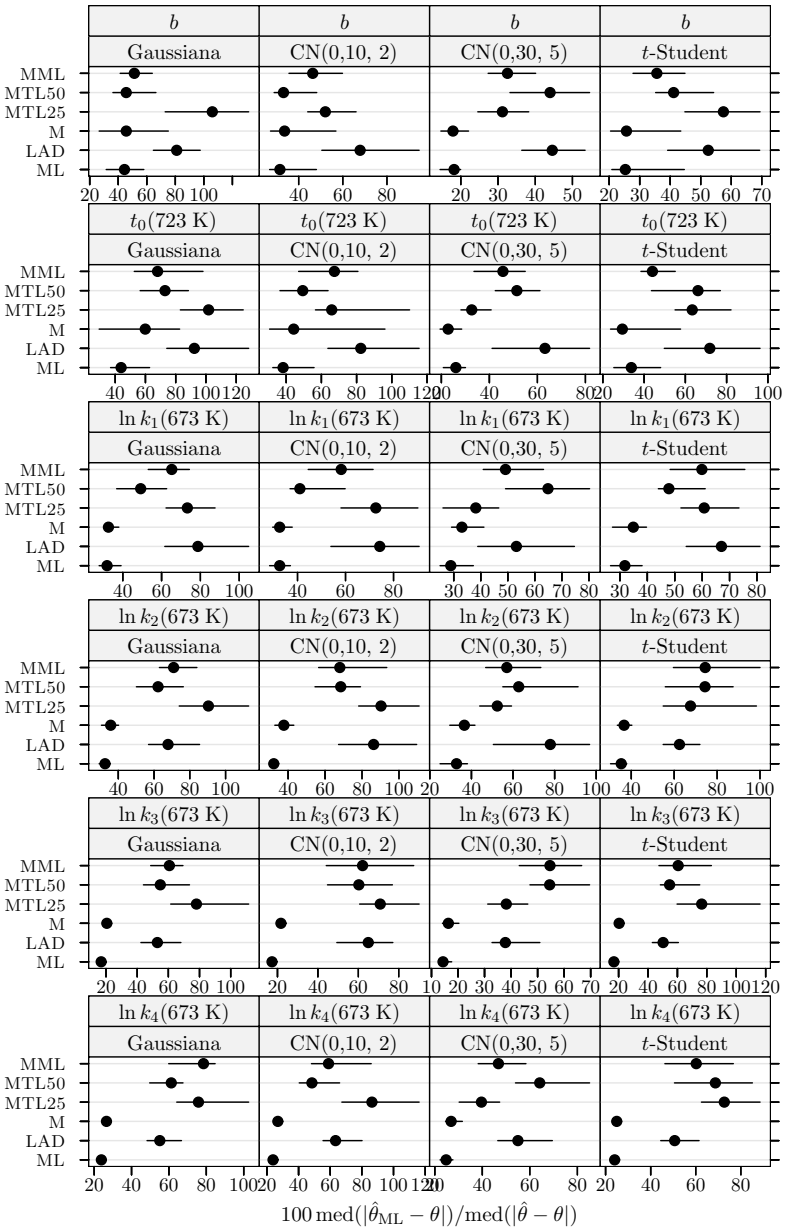


Figura 5.25 Pirólise do xisto betuminoso: medida da eficiência dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 10, 19, 20, 24, 32, 36, 37, 38, 42, 58, e 63.

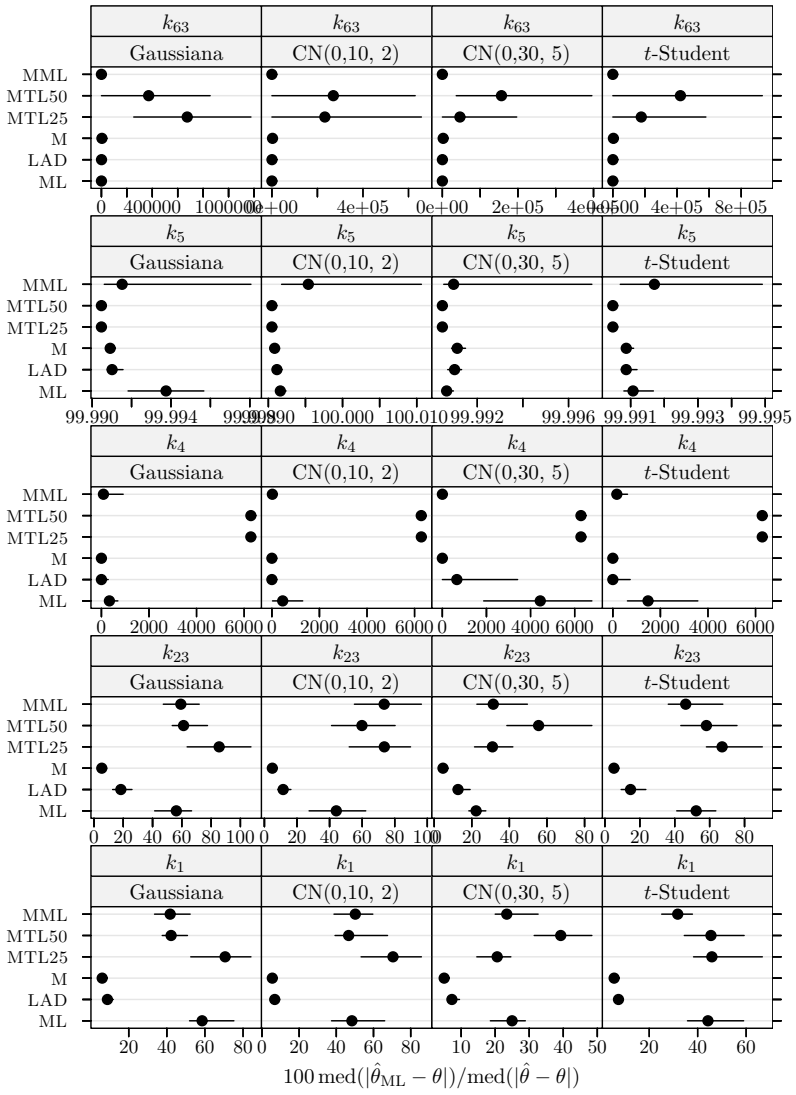


Figura 5.26 Conversão do metanol em hidrocarbonetos: medida da eficiência dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, e 9.

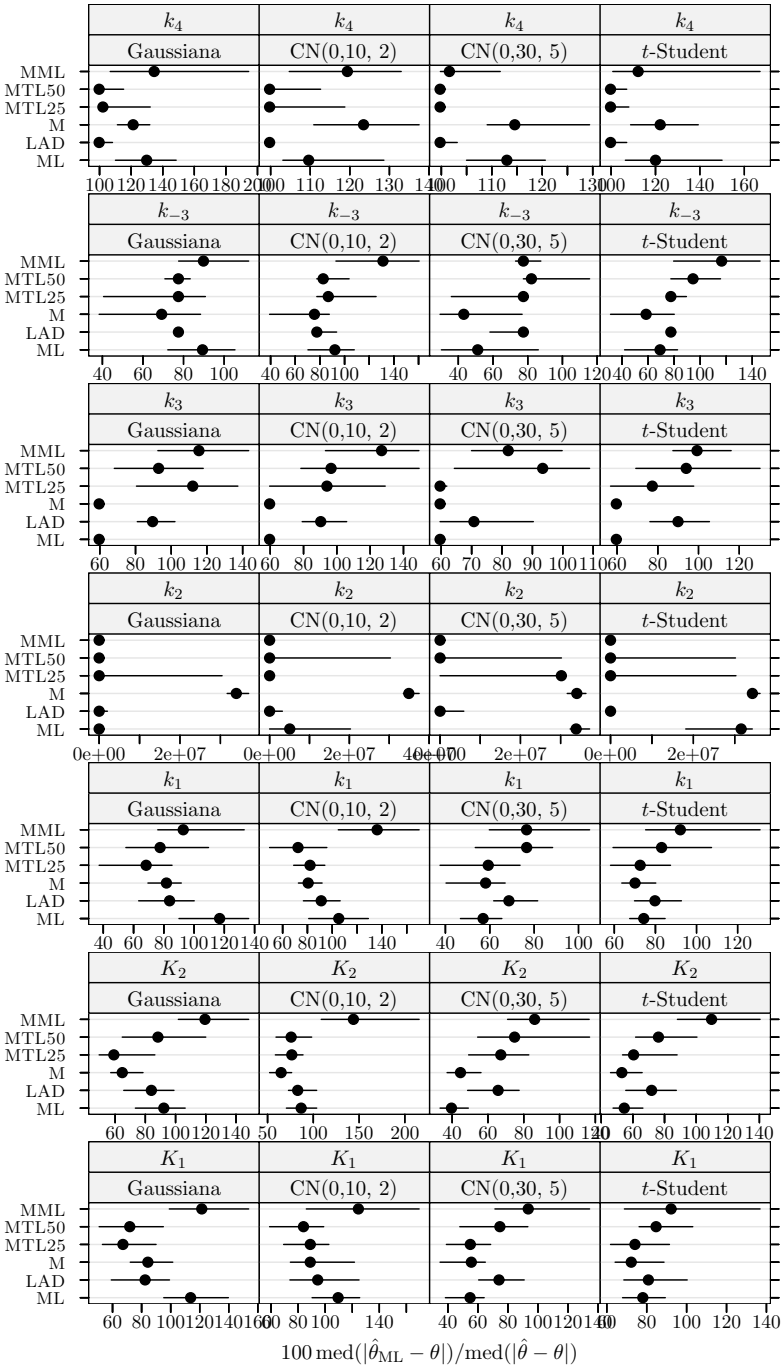


Figura 5.27 Hidrogenação catalítica do 3-hidroxipropanal: medida da eficiência dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 7, 11, 22, 23, 27, 29, e 30.

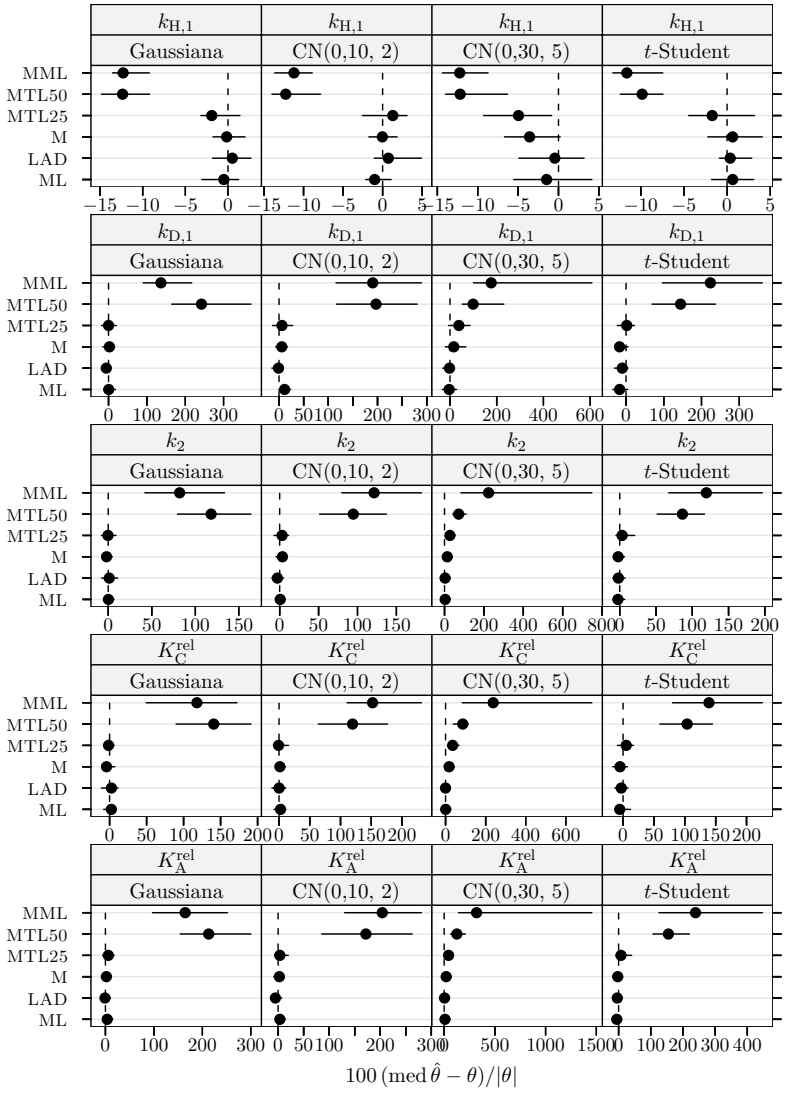


Figura 5.28 Hidrogenação do tolueno: índice de enviesamento robustificado dos estimadores para dados simulados com 20% de *outliers* e $\delta_R = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, e 6.

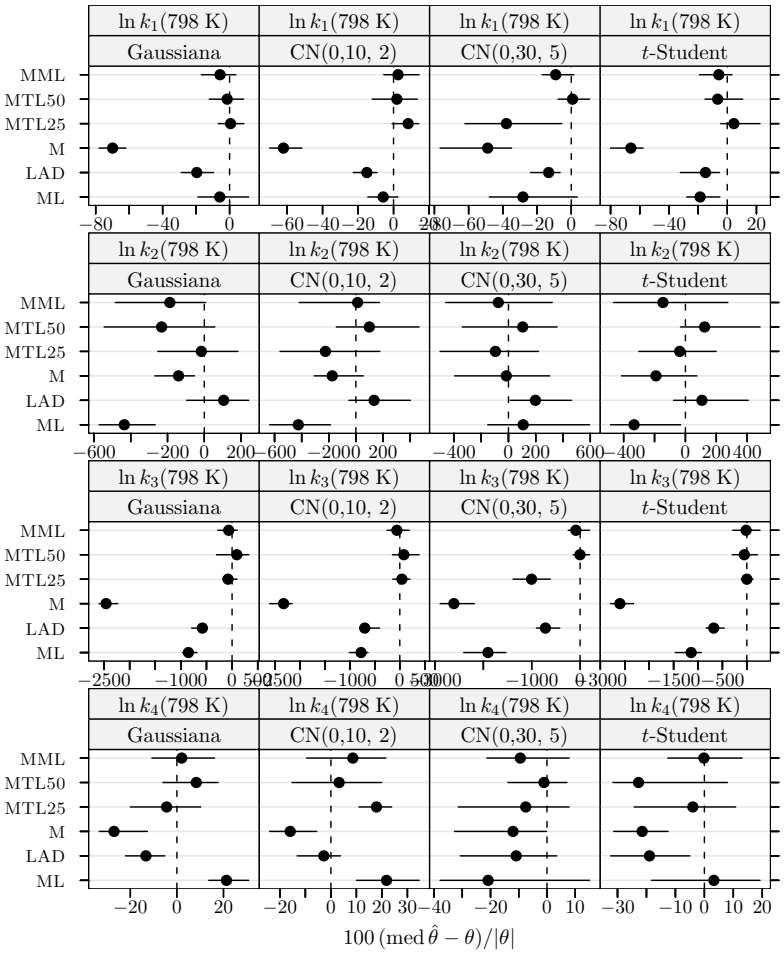


Figura 5.29 Pirólise do xisto betuminoso: índice de enviesamento robustificado dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 10, 19, 20, 24, 32, 36, 37, 38, 42, 58, e 63.

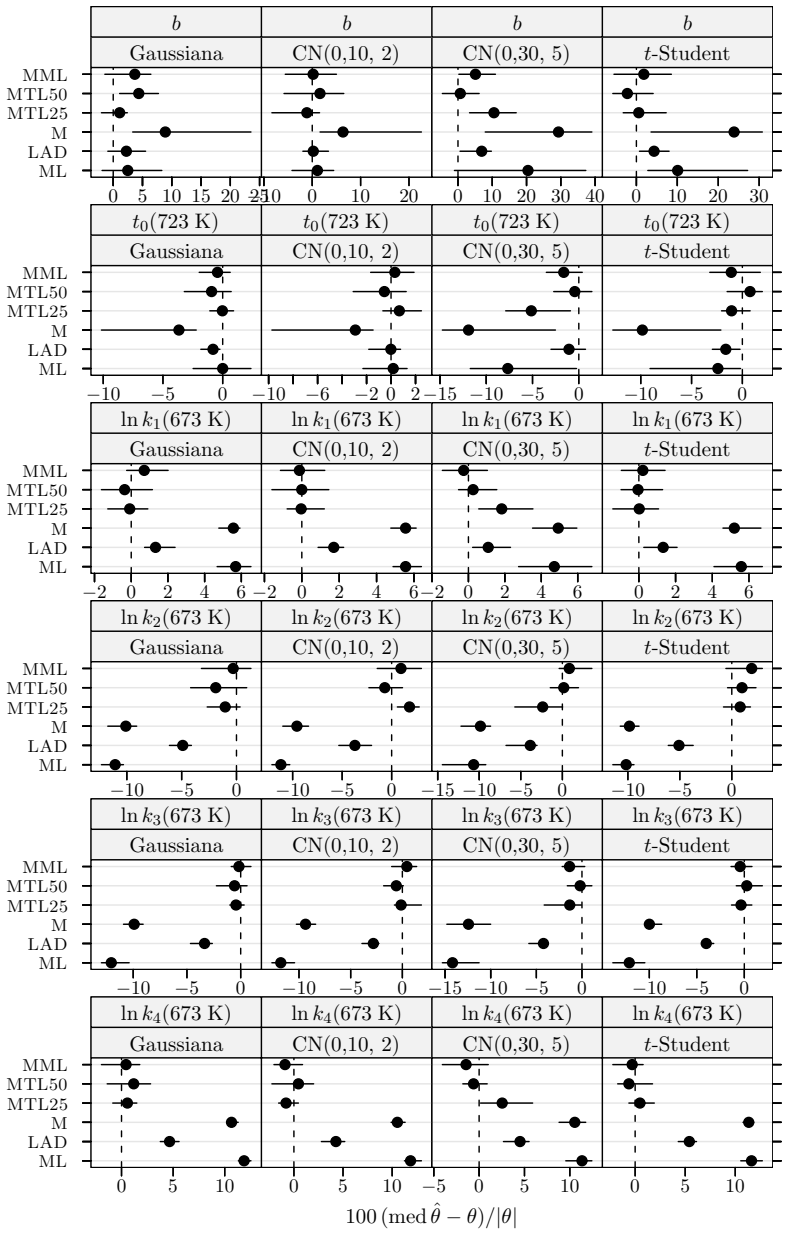


Figura 5.30 Pirólise do xisto betuminoso: índice de enviesamento robustificado dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 10, 19, 20, 24, 32, 36, 37, 38, 42, 58, e 63.

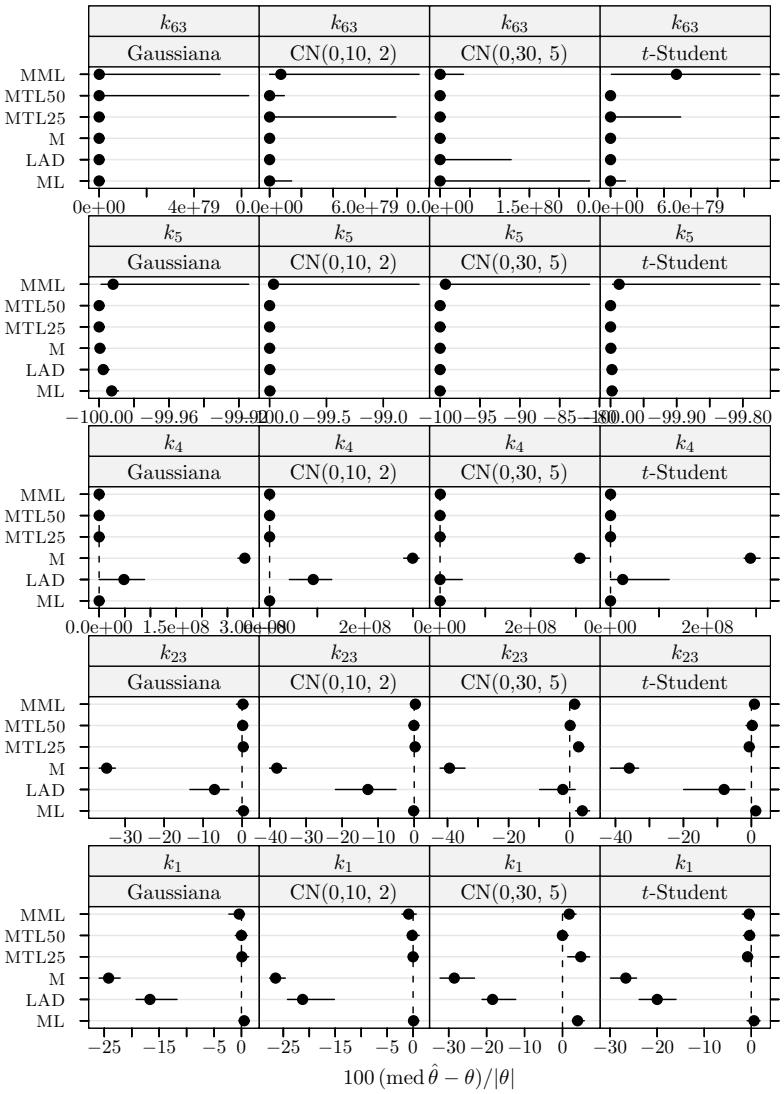


Figura 5.31 Conversão do metanol em hidrocarbonetos: índice de enviesamento robustificado dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, e 9.

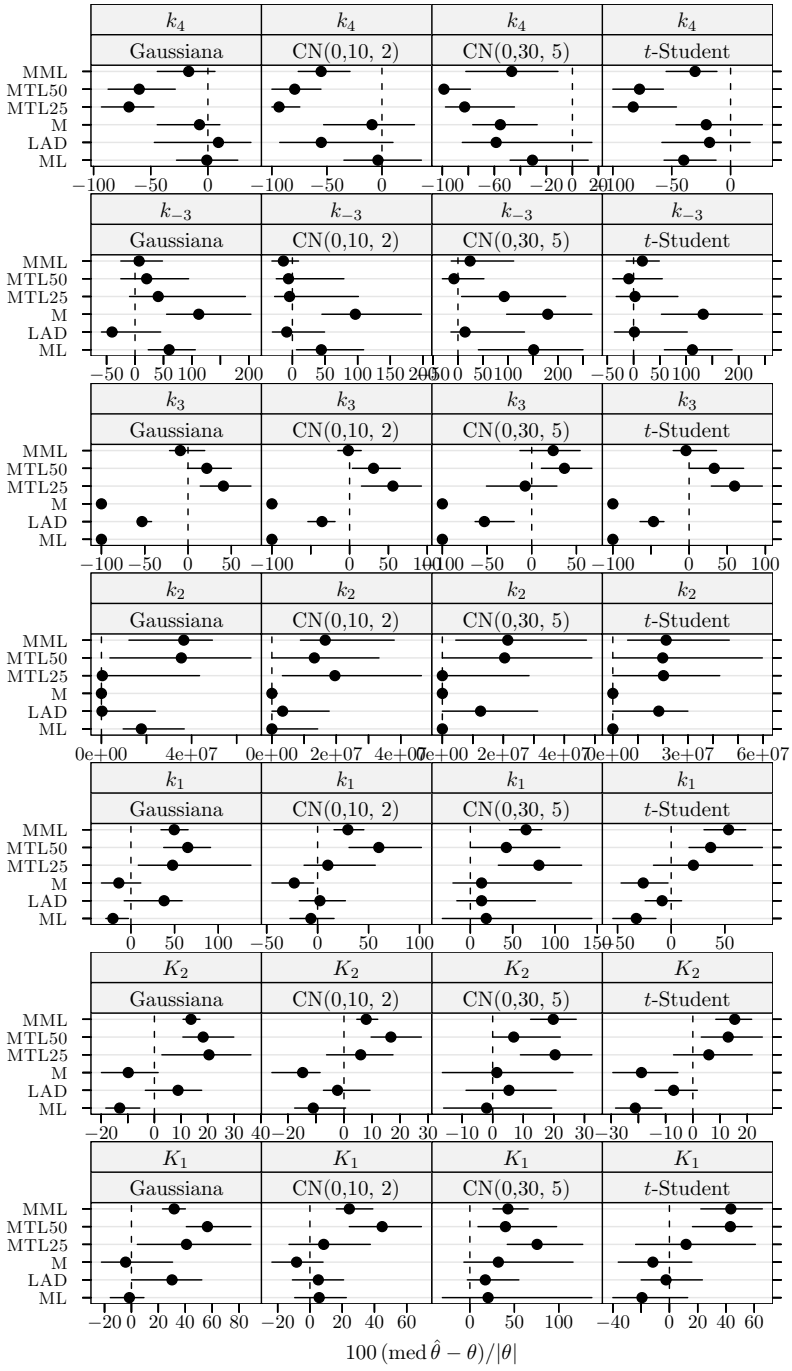


Figura 5.32 Hidrogenação catalítica do 3-hidroxipropanal: índice de enviesamento robustificado dos estimadores para dados simulados com 20% de *outliers* e $\delta_{\mathrm{R}} = 5$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 7, 11, 22, 23, 27, 29, e 30.

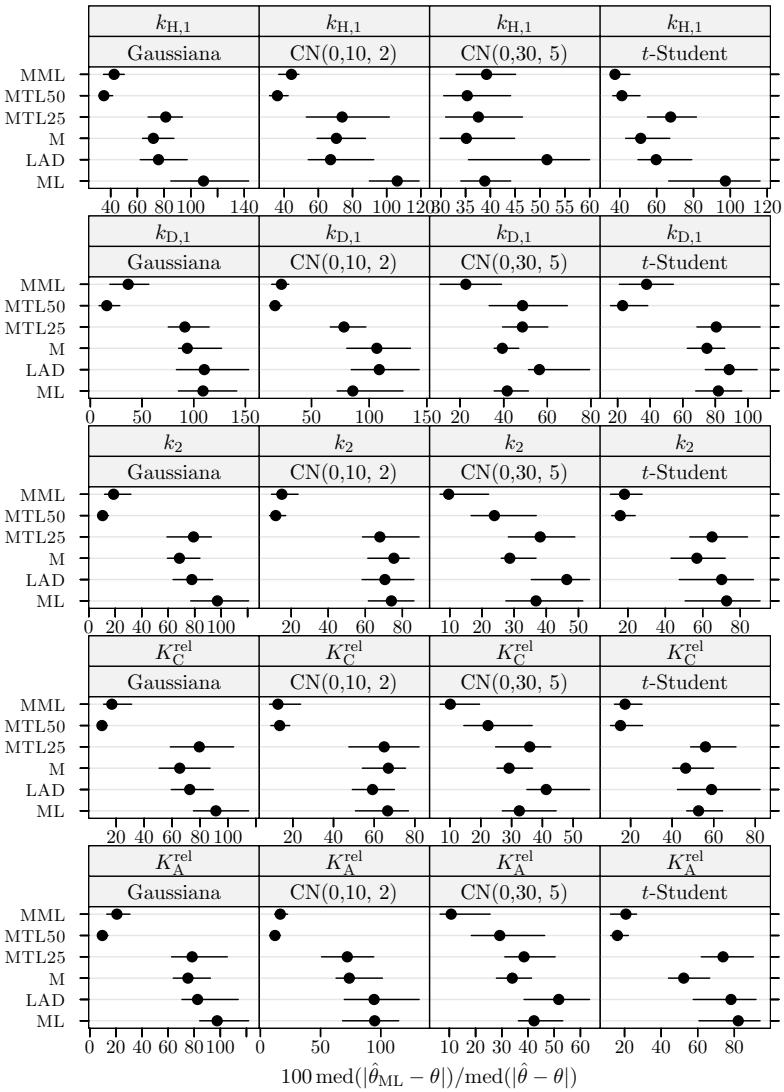


Figura 5.33 Hidrogenação do tolueno: medida da eficiência dos estimadores para dados simulados com 15% de *outliers* e $\delta_R = 10$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1 e 6.

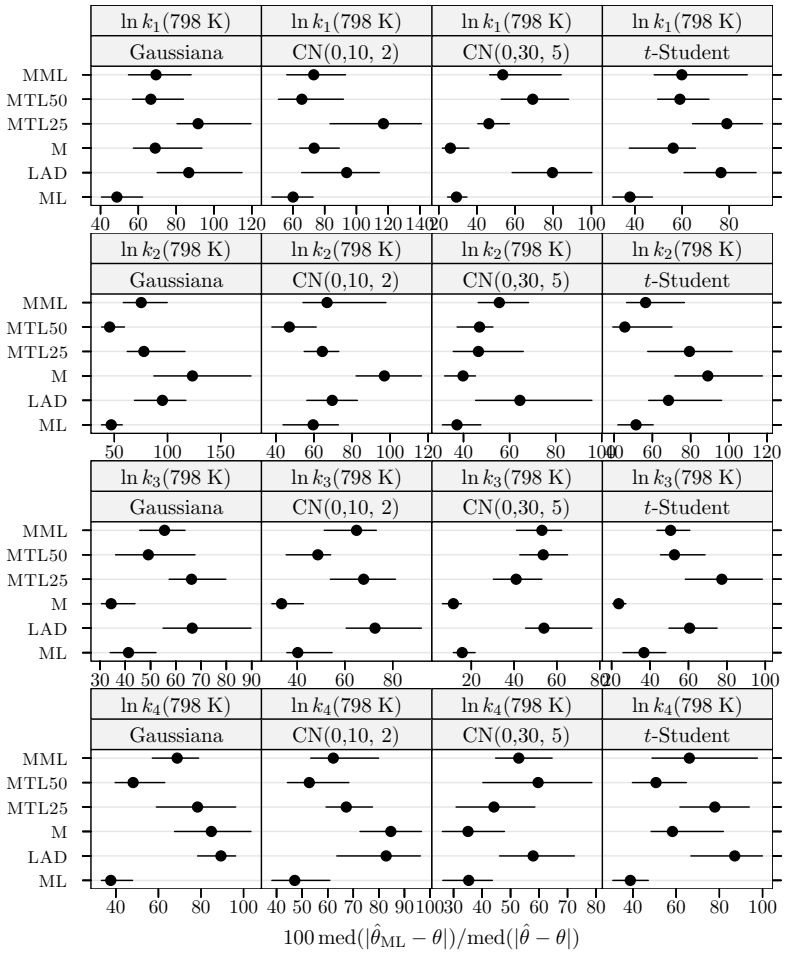


Figura 5.34 Pirólise do xisto betuminoso: medida da eficiência dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 10, 19, 32, 36, 37, 38, 42, e 58.

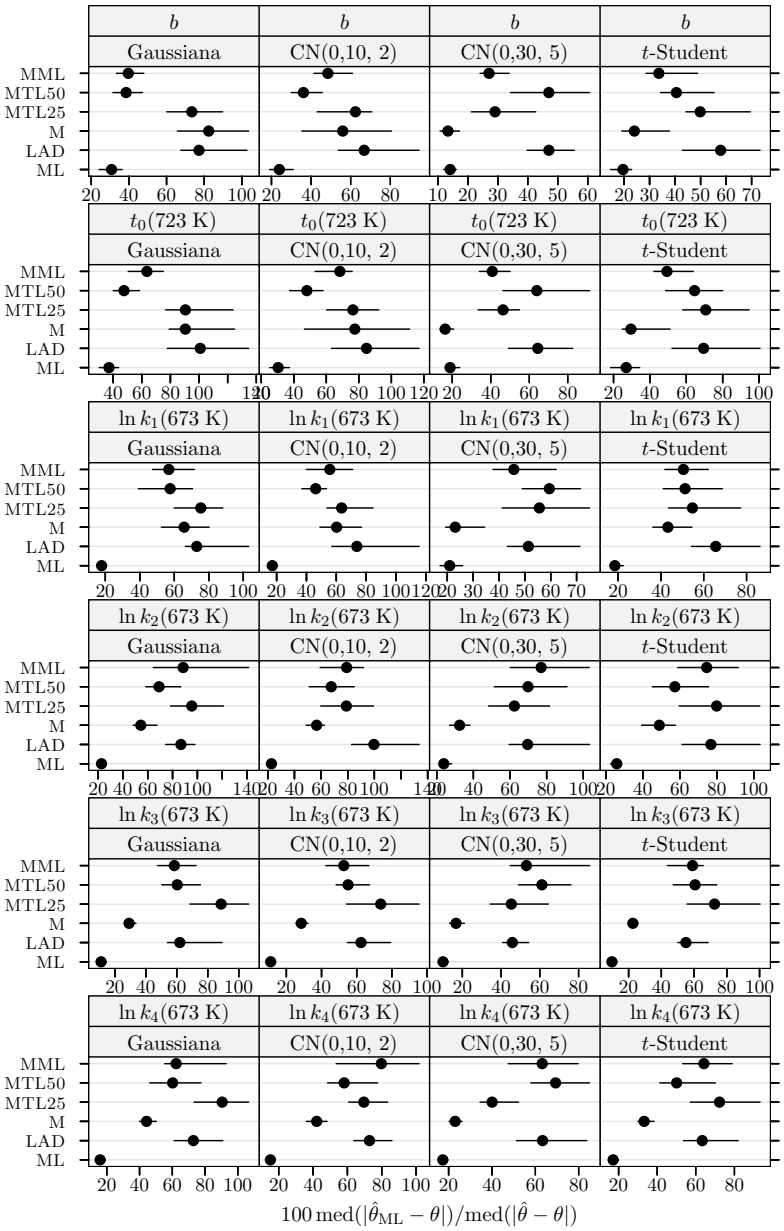


Figura 5.35 Pirólise do xisto betuminoso: medida da eficiência dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 10, 19, 32, 36, 37, 38, 42, e 58.

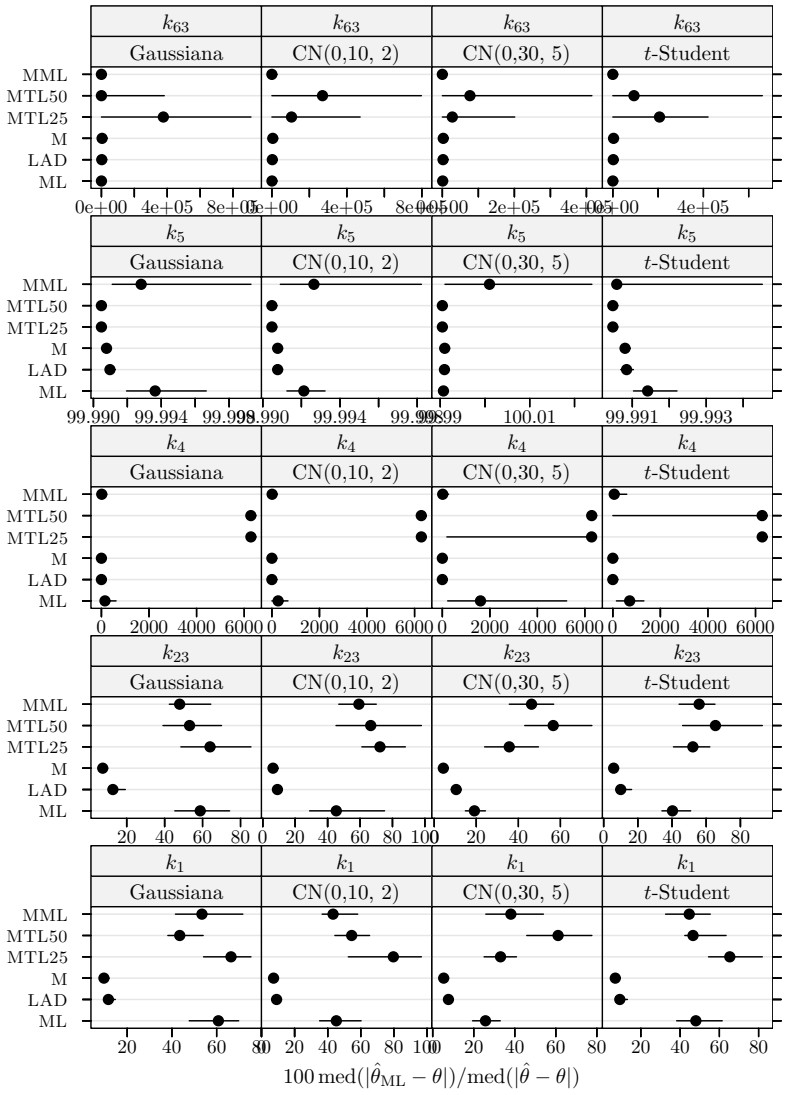


Figura 5.36 Conversão do metanol em hidrocarbonetos: medida da eficiência dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 6 e 8.

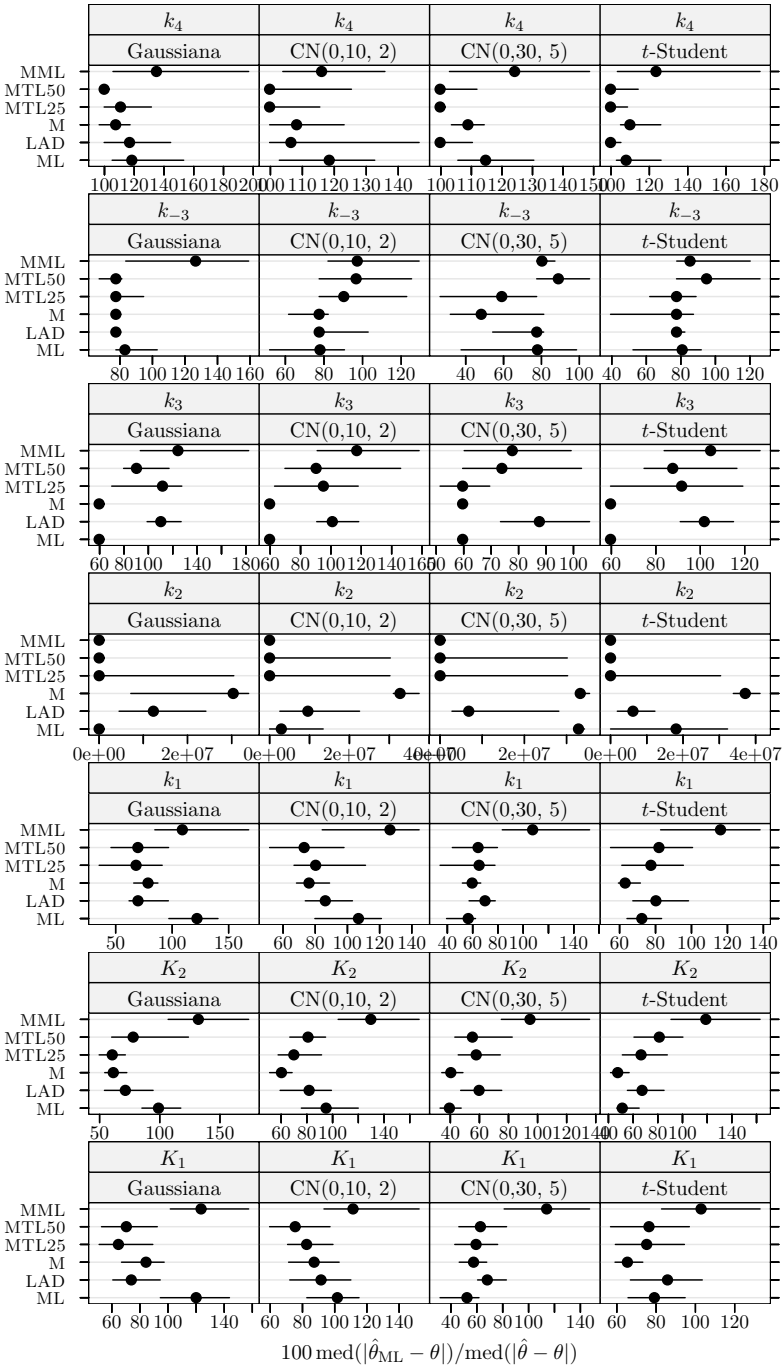


Figura 5.37 Hidrogenação catalítica do 3-hidroxipropanal: medida da eficiência dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 11, 22, 23, 27, 29, e 30.

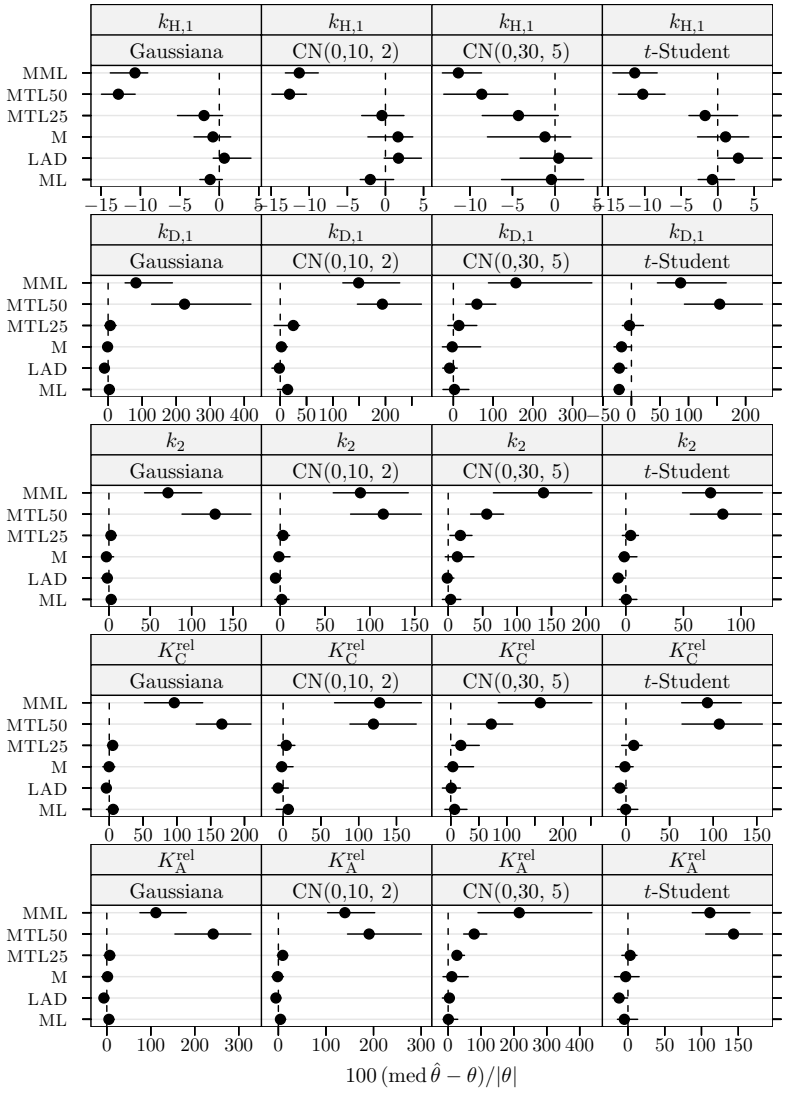


Figura 5.38 Hidrogenação do tolueno: índice de enviesamento robustificado dos estimadores para dados simulados com 15% de *outliers* e $\delta_R = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1 e 6.

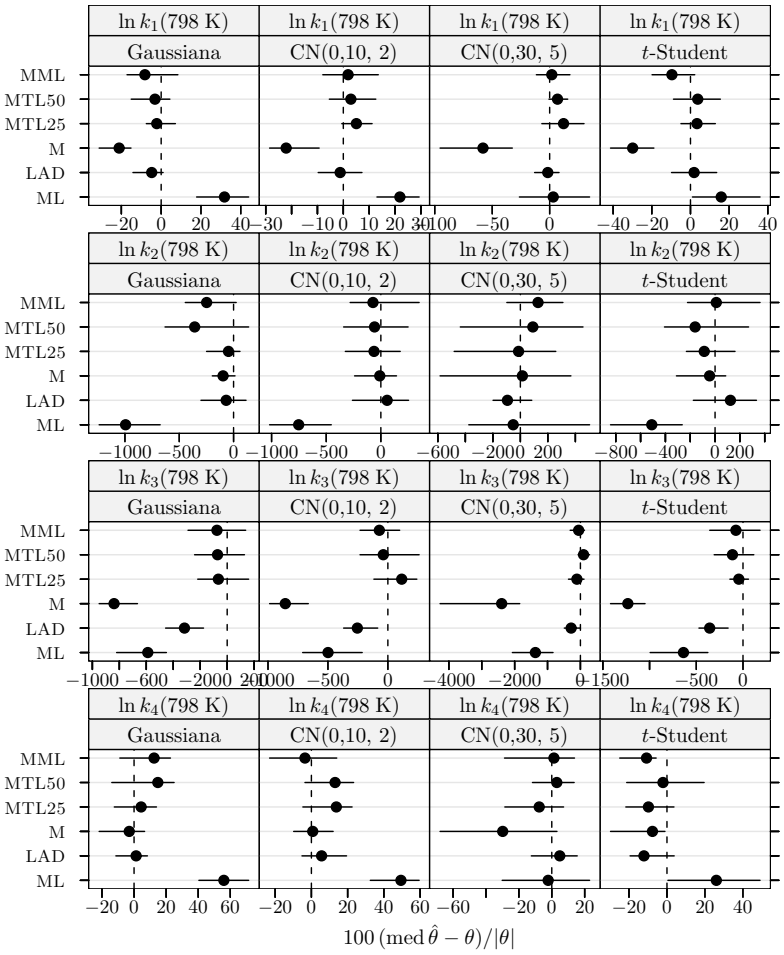


Figura 5.39 Pirólise do xisto betuminoso: índice de enviesamento robustificado dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 10, 19, 32, 36, 37, 38, 42, e 58.

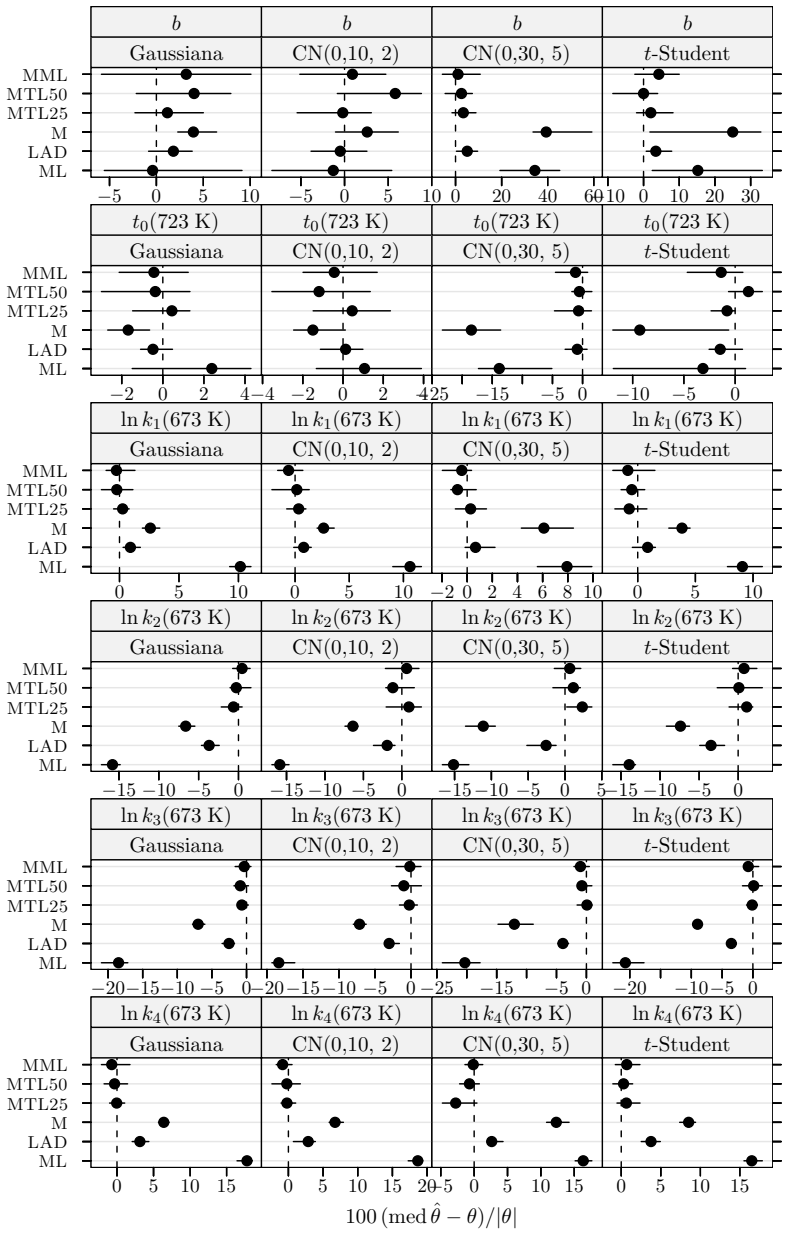


Figura 5.40 Pirólise do xisto betuminoso: índice de enviesamento robustificado dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 10, 19, 32, 36, 37, 38, 42, e 58.

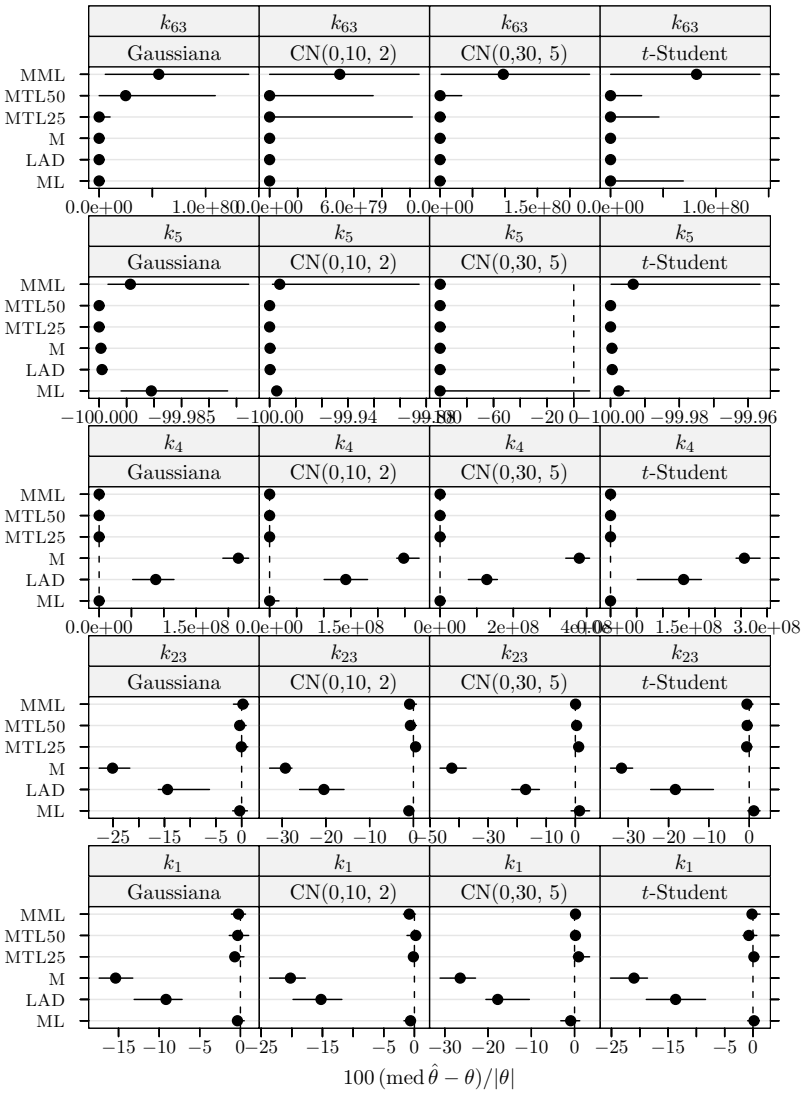


Figura 5.41 Conversão do metanol em hidrocarbonetos: índice de enviesamento robustificado dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 6 e 8.

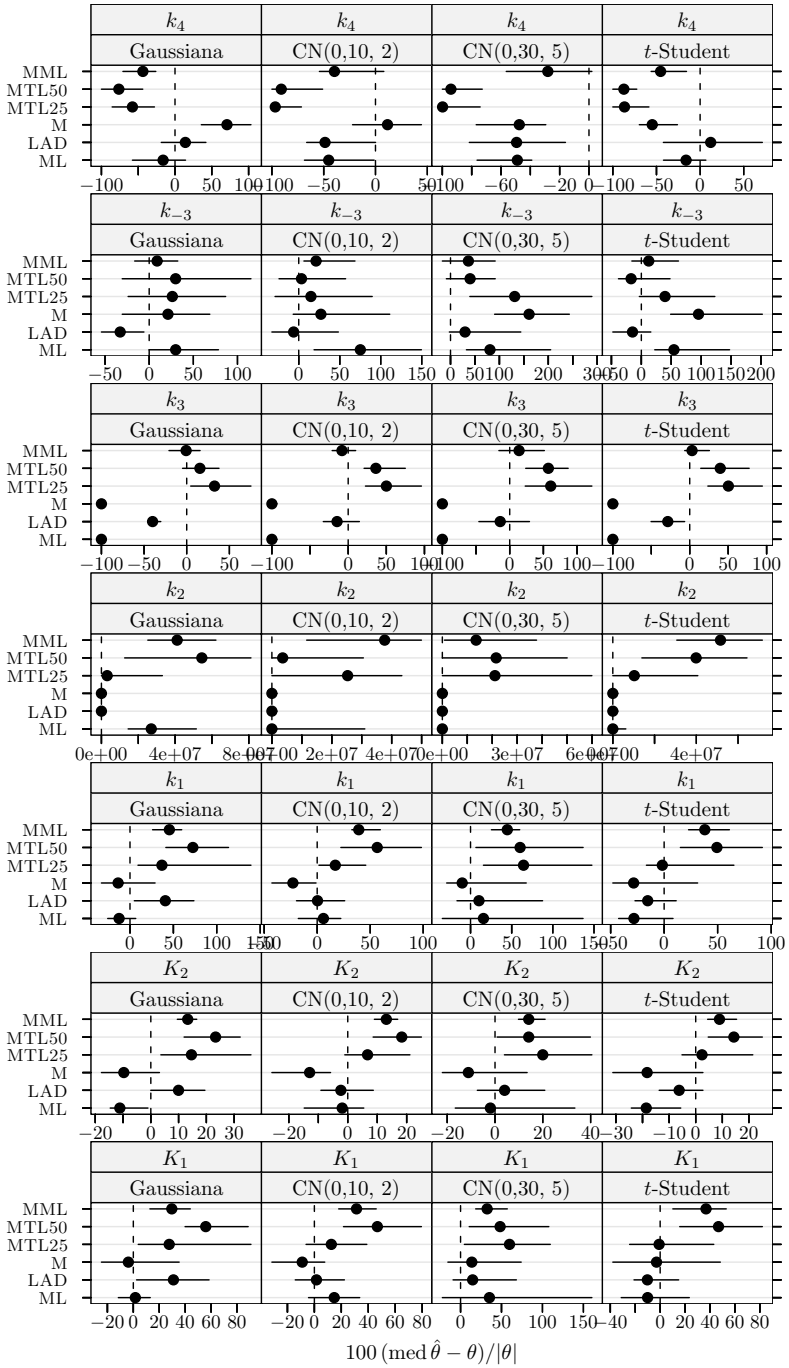


Figura 5.42 Hidrogenação catalítica do 3-hidroxipropanal: índice de enviesamento robustificado dos estimadores para dados simulados com 15% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 11, 22, 23, 27, 29, e 30.

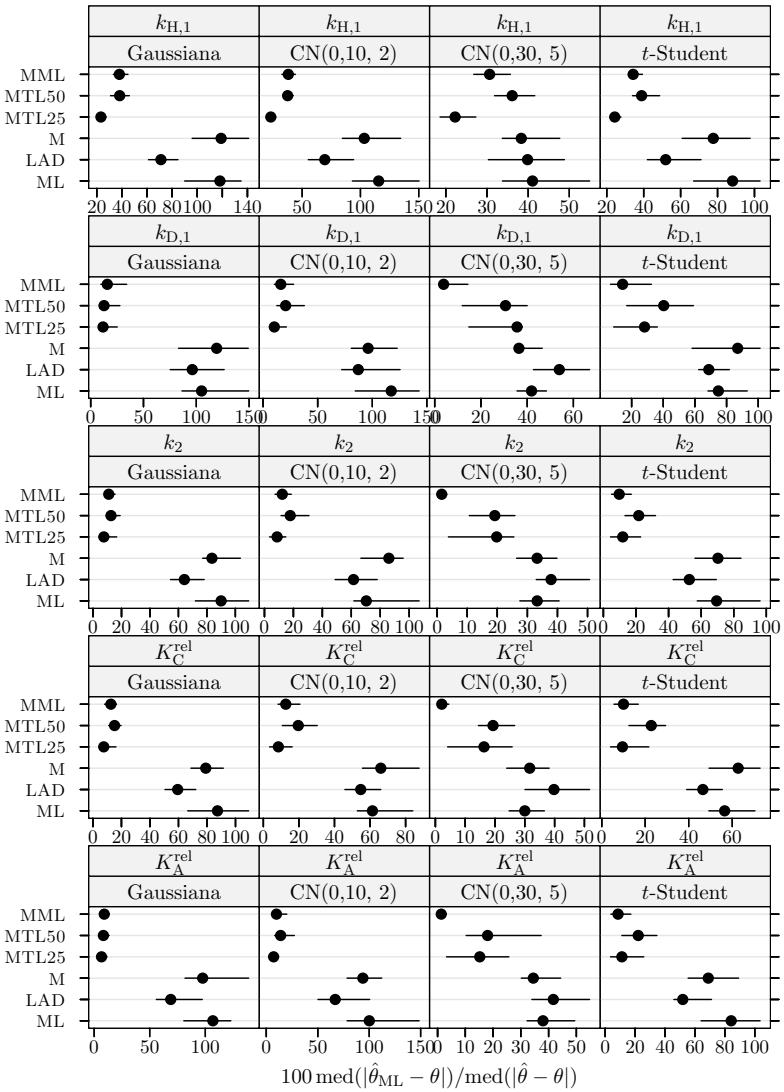


Figura 5.43 Hidrogenação do tolueno: medida da eficiência dos estimadores para dados simulados com 30% de *outliers* e $\delta_R = 10$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 5, e 6.

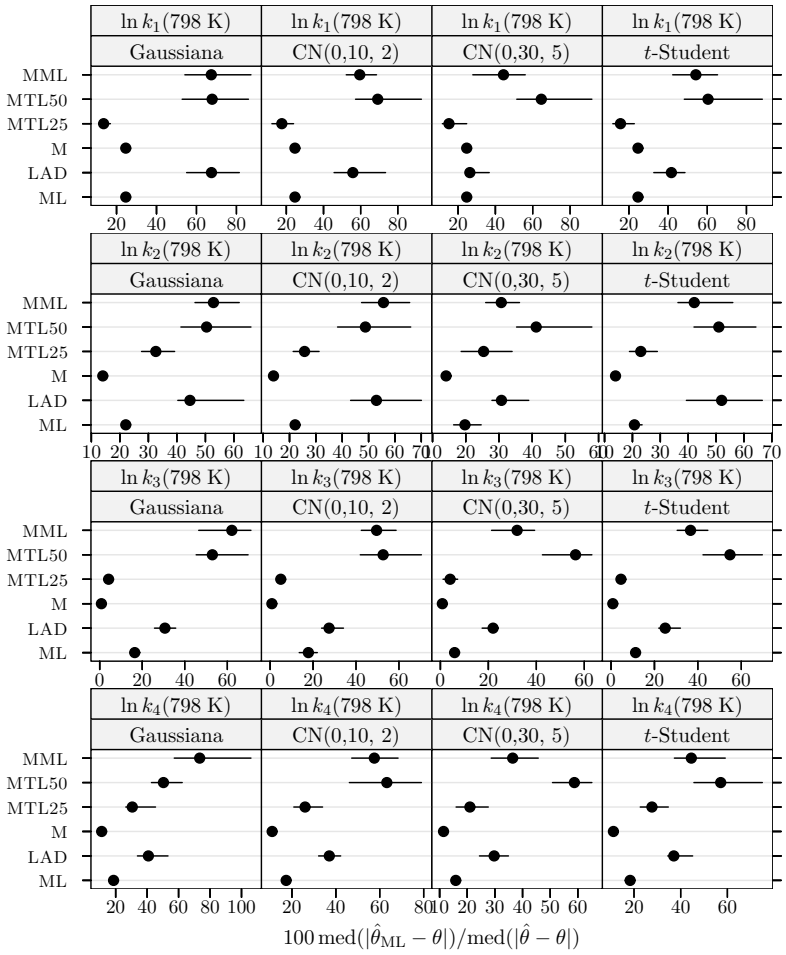


Figura 5.44 Pirólise do xisto betuminoso: medida da eficiência dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 10, 19, 20, 22, 24, 26, 32, 36, 37, 38, 42, 44, 53, 58, 60, 62, e 63.

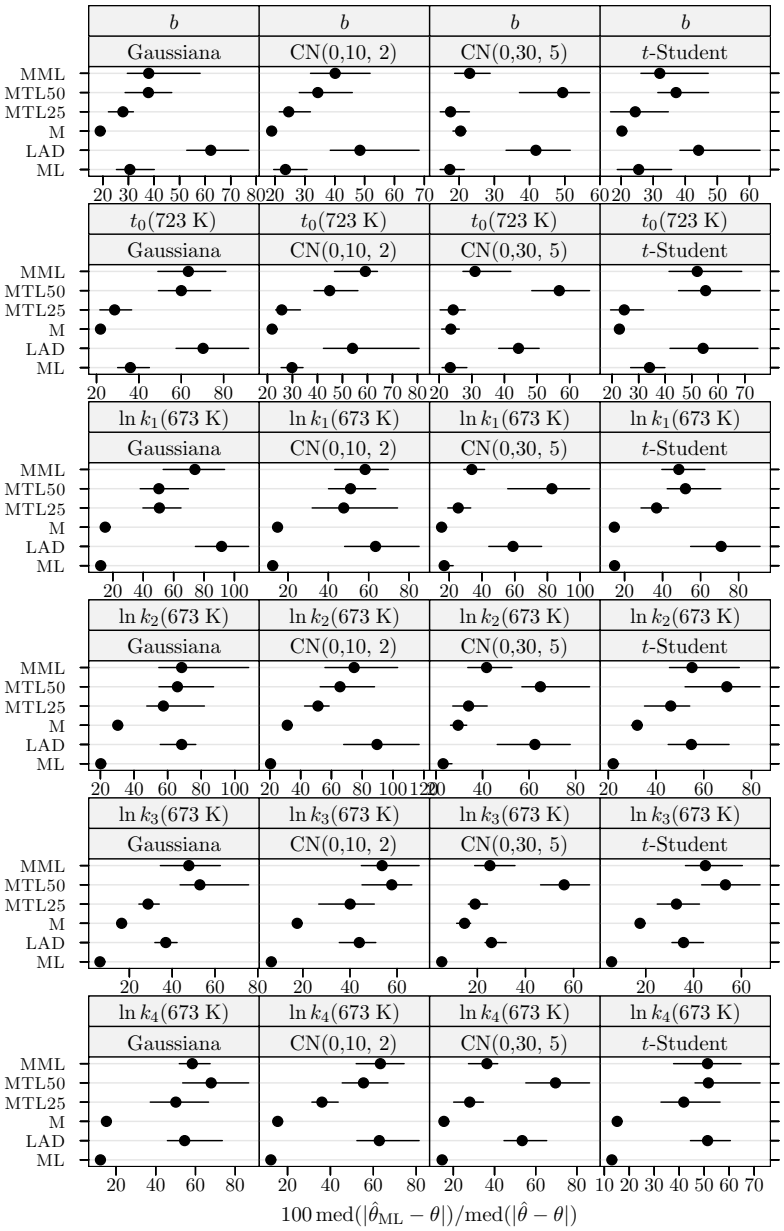


Figura 5.45 Pirólise do xisto betuminoso: medida da eficiência dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 10, 19, 20, 22, 24, 26, 32, 36, 37, 38, 42, 44, 53, 58, 60, 62, e 63.

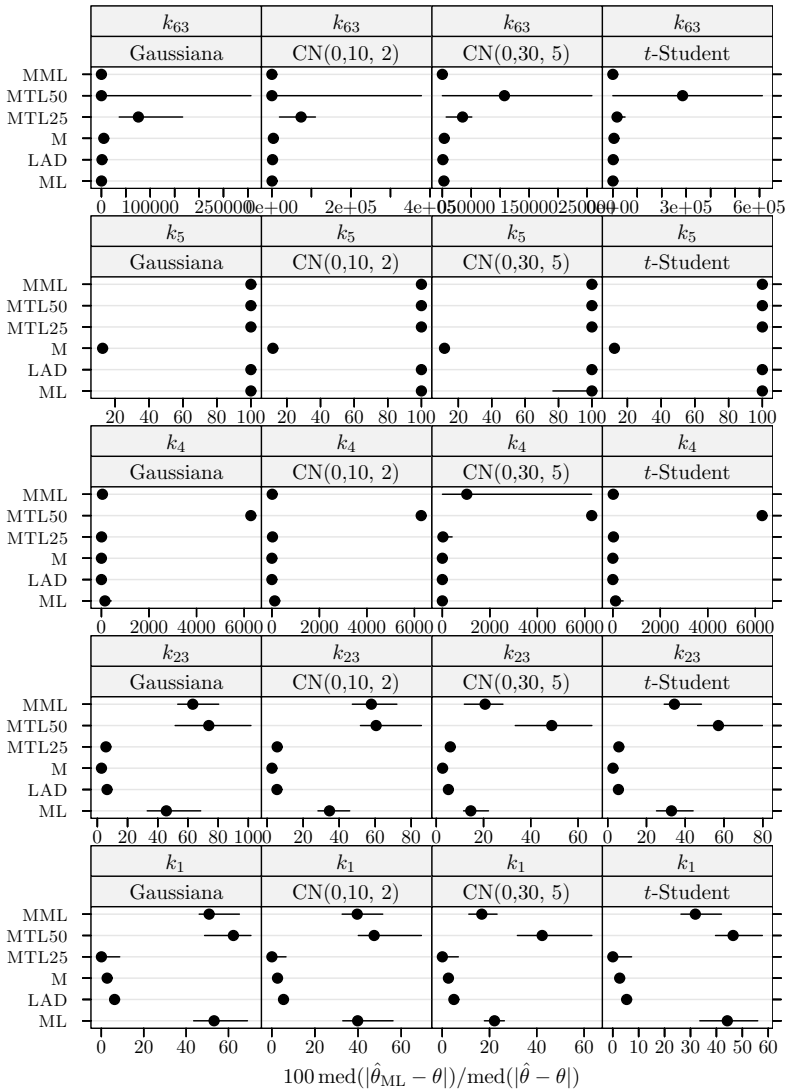


Figura 5.46 Conversão do metanol em hidrocarbonetos: medida da eficiência dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 3, 8, 9, 13, e 16.

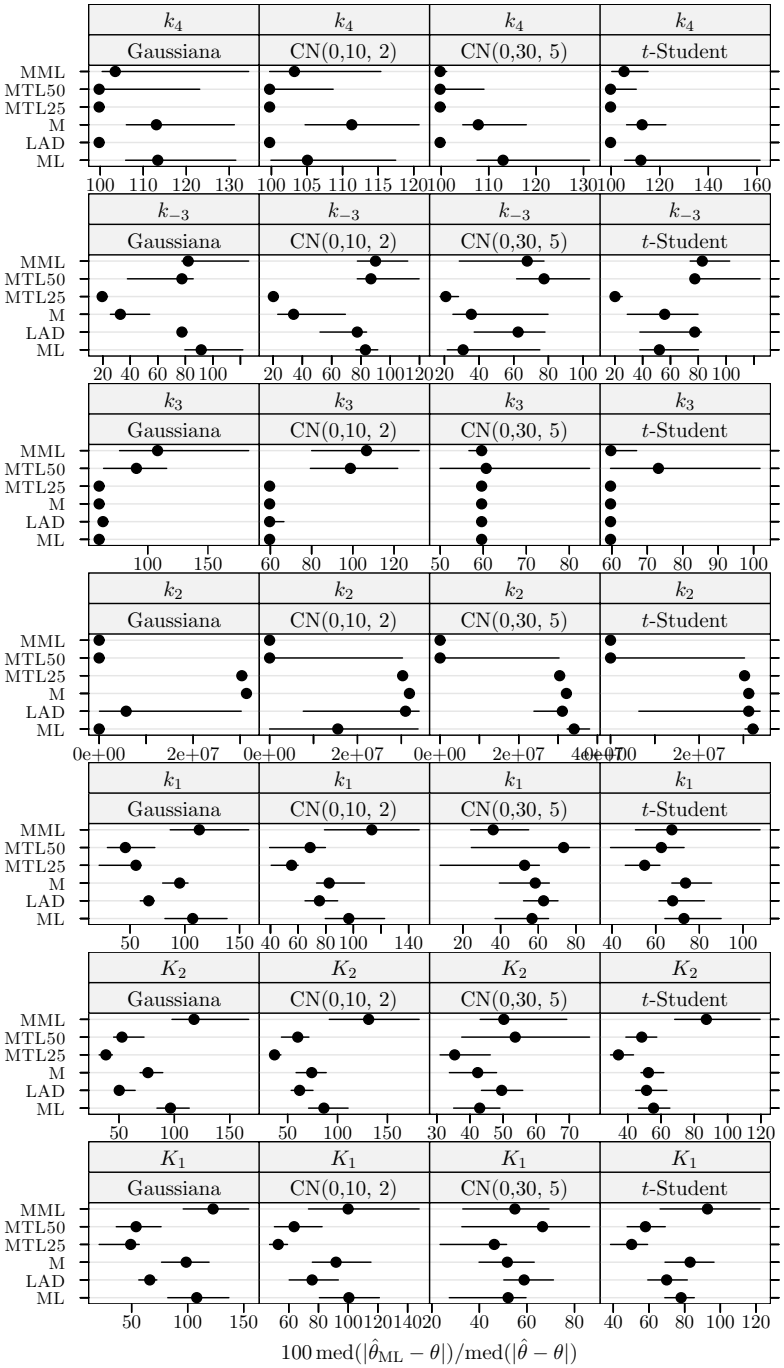


Figura 5.47 Hidrogenação catalítica do 3-hidroxipropanal: medida da eficiência dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. A eficiência é relativa ao estimador do critério do determinante ajustado a observações apenas com erro Gaussiano. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 7, 8, 10, 11, 13, 22, 23, 27, 28, 29, e 30.

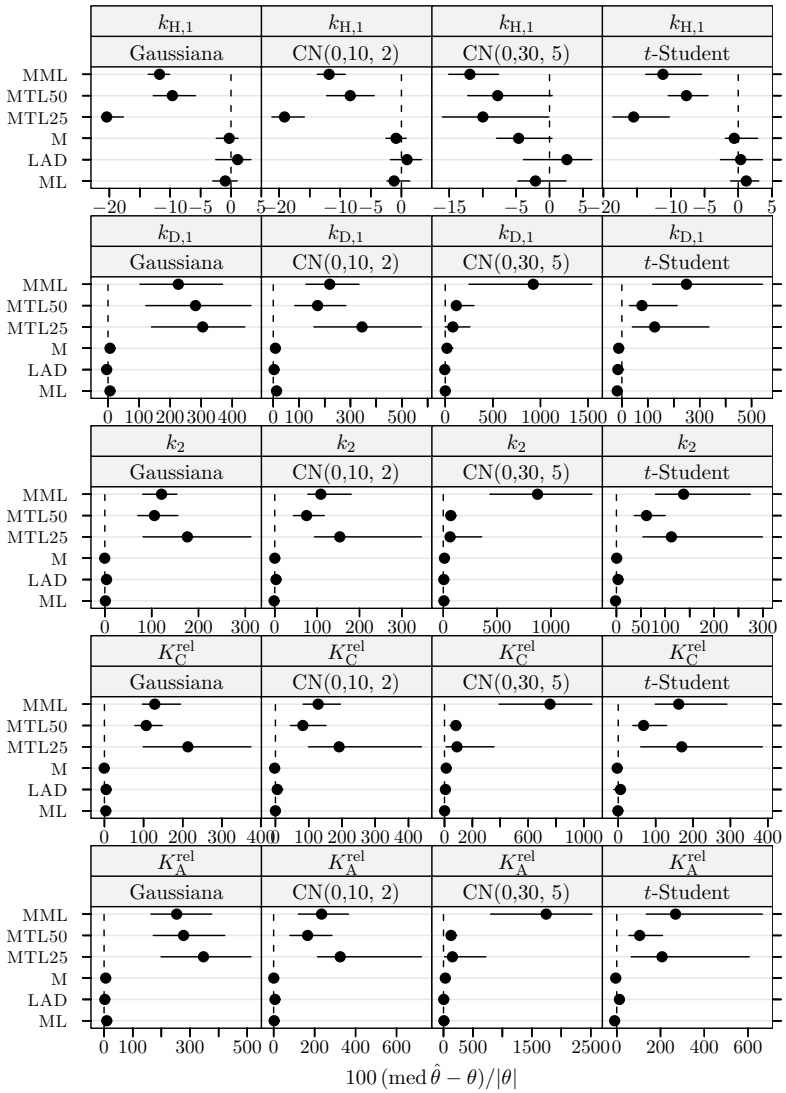


Figura 5.48 Hidrogenação do tolueno: índice de enviesamento robustificado dos estimadores para dados simulados com 30% de *outliers* e $\delta_R = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 5, e 6.

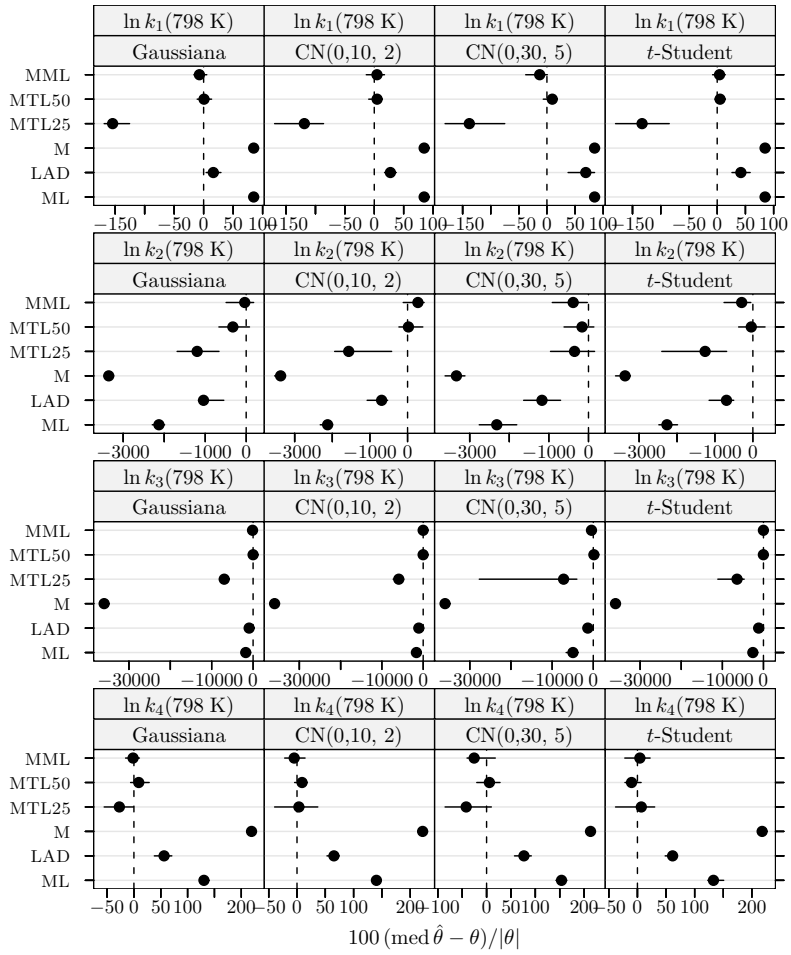


Figura 5.49 Pirólise do xisto betuminoso: índice de enviesamento robustificado dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 10, 19, 20, 22, 24, 26, 32, 36, 37, 38, 42, 44, 53, 58, 60, 62, e 63.

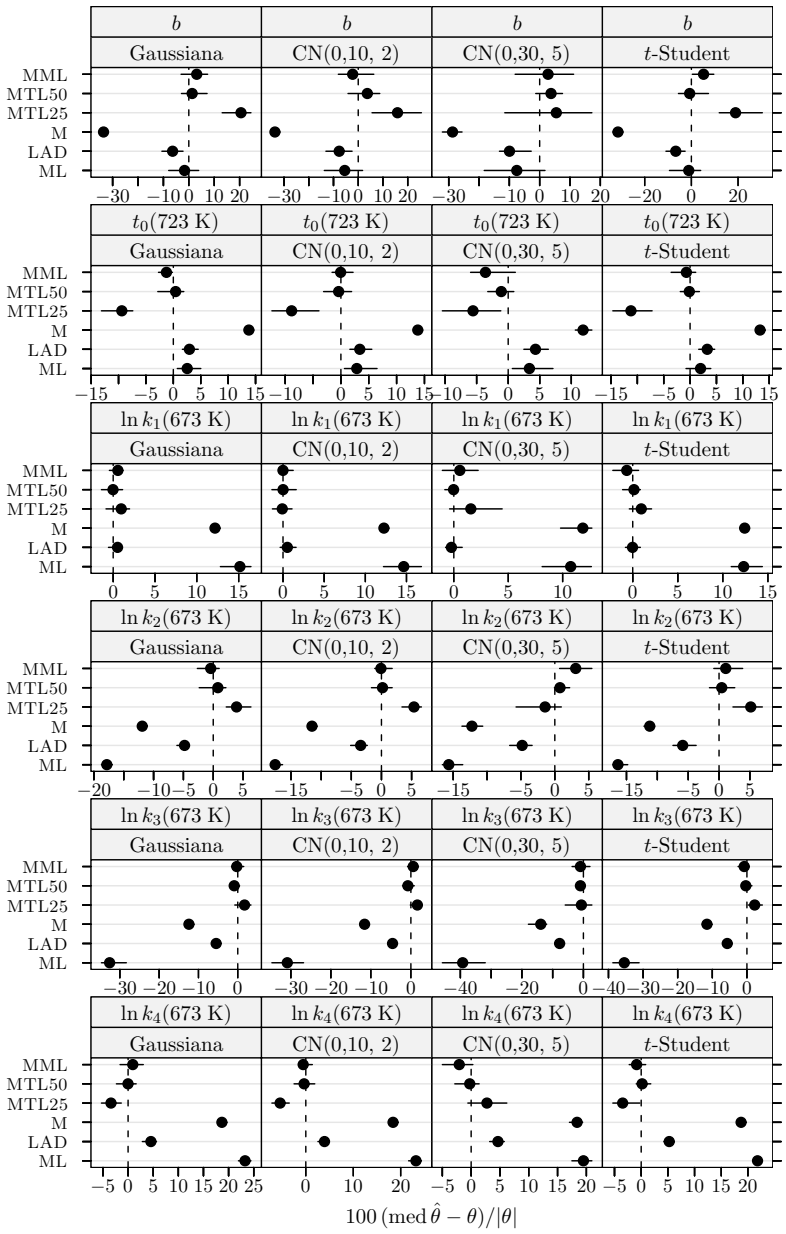


Figura 5.50 Pirólise do xisto betuminoso: índice de enviesamento robustificado dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 1, 4, 10, 19, 20, 22, 24, 26, 32, 36, 37, 38, 42, 44, 53, 58, 60, 62, e 63.

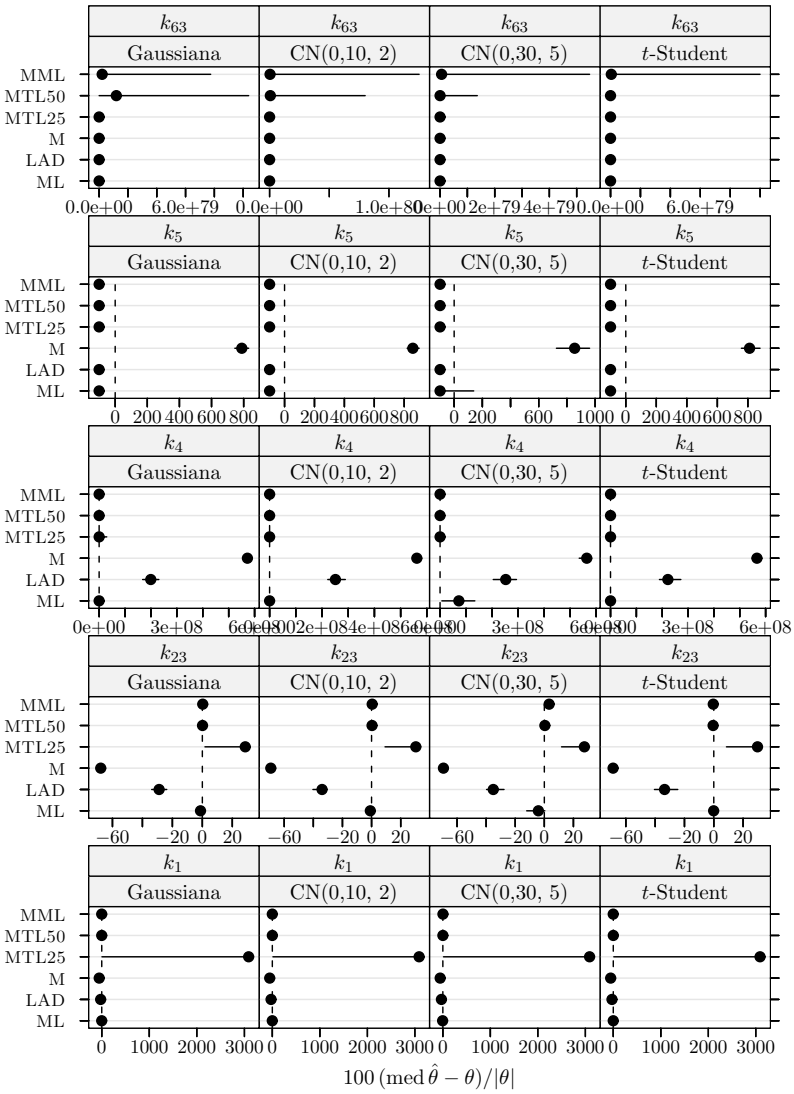


Figura 5.51 Conversão do metanol em hidrocarbonetos: índice de enviesamento robustificado dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 3, 8, 9, 13, e 16.

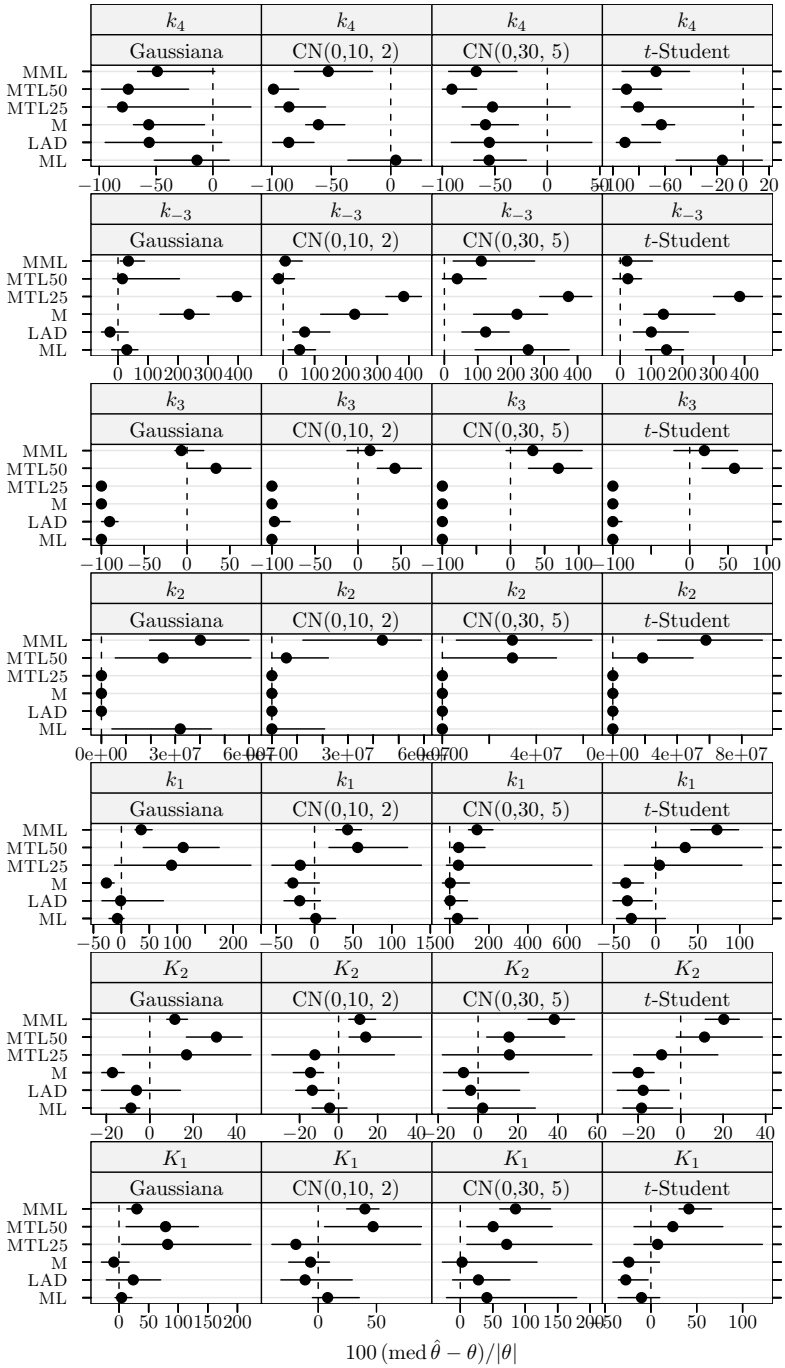


Figura 5.52 Hidrogenação catalítica do 3-hidroxipropanal: índice de enviesamento robustificado dos estimadores para dados simulados com 30% de *outliers* e $\delta_{\mathrm{R}} = 10$. Cada ponto representa o valor do índice, enquanto o segmento de recta a cheio representa o intervalo de confiança a 95% calculado usando o método do percentil da técnica *bootstrap* com 999 amostras. Os pontos perturbados foram 7, 8, 10, 11, 13, 22, 23, 27, 28, 29, e 30.

5.7 Discussão das experiências com dados simulados com outliers

Caso de contaminação moderada (10% de outliers, $\delta_R = 5$)

Na mesma linha do cenário anterior, também aqui a análise dos resultados compilados nas figuras 5.13 a 5.17 nas páginas 137–141 permite revelar que o impacto de *outliers* nos diferentes problemas estudados foi de sentidos diversos.

De facto, no caso da hidrogenação do tolueno, observe-se a coincidência (em termos relativos) da eficiência quando se passa do cenário sem *outliers* para o cenário presente. Contudo, analisando os casos da pirólise do xisto betuminoso e da conversão do metanol em hidrocarbonetos constata-se que há uma perda de eficiência elucidativa do estimador do critério do determinante em relação aos melhores estimadores robustos. Relativamente à hierarquização do desempenho dos vários estimadores robustos, continua a observar-se o descrito na secção anterior. Novamente no caso da hidrogenação catalítica do 3-hidroxipropanal não pode ir-se muito longe na identificação de um conjunto de características distintas, à excepção da superioridade do estimador MML.

Comparando com a situação sem *outliers* é interessante notar o aumento considerável do enviesamento dos estimadores LAD e M para o caso da conversão do metanol em hidrocarbonetos.

Caso de contaminação severa (20% de outliers, $\delta_R = 5$)

Como se pode observar pelas figuras 5.23 a 5.32 nas páginas 147–156 as características principais do cenário vertente são basicamente as mesmas do cenário anterior com excepção do caso da pirólise do xisto betuminoso, no qual o estimadores MTL50 e LAD são relativamente melhores que os restantes estimadores robustos, e o estimador M tem um comportamento bastante pobre.

Caso de forte contaminação e perturbação (15% de outliers, $\delta_R = 10$)

Os resultados obtidos evidenciam uma situação de quase redundância em relação ao cenário anterior.

Caso limite (30% de outliers, $\delta_R = 10$)

Também aqui os resultados das figuras 5.43 a 5.47 nas páginas 167–171 são análogos aos do cenário de contaminação severa. A diferença básica entre entre o caso presente e o caso de contaminação severa reside no decaimento brusco do desempenho do estimador MTL25, aliás como seria de esperar já que a fracção de contaminação dos dados por *outliers* é superior à fracção de aparamento do estimador. Verifica-se, à excepção do caso da hidrogenação do tolueno, que os melhores estimadores são estimadores MML e MTL50.

5.8 Comentários finais

Contrariamente ao caso de resposta univariada, no caso multivariado há muitas situações de estimadores com comportamento que varia de problema para problema. Como exemplo disso, considere-se os estimadores M e LAD no problema da hidrogenação do tolueno contra o problema da conversão do metanol em hidrocarbonetos. Observe-se ainda como nos cenários com *outliers* do problema da hidrogenação do tolueno o estimador do critério do determinante partilha o melhor desempenho entre os estimadores em competição, enquanto nos restantes problemas tal não sucede. Sendo assim, aqui torna-se problemático esquematizar o comportamento dos diferentes estimadores.

Evidentemente, dada a insuficiência do número de problemas analisados, não é possível ajuizar se esta característica é relativamente universal ou, ao contrário, algo exótica.

Capítulo 6

Conclusões

6.1 Observações gerais

Perante a diversidade dos métodos de regressão robustos propostos na literatura científica com propriedades de amostra finita desconhecidas, surge a questão de saber que alternativa usar.

Neste sentido, foi realizado um estudo de Monte Carlo para investigar o desempenho de estimadores robustos, assim como para uma comparação com o estimador "clássico". O caso de modelos não-lineares com uma só resposta integra os métodos LS, L_p, LMS, LTS, MM, τ, e LTD. Os métodos incluídos no caso de resposta multivariada foram: ML (critério do determinante), LAD, MML, e MTL.

Deve sublinhar-se que, tanto quanto se sabe, é realizado pela primeira vez neste trabalho (a) o estudo integrado destes estimadores no domínio da Estatística, e de qualquer deles no domínio da Engenharia Química e (b) são aplicados pela primeira vez ao caso multivariado os estimadores MML, e MTL.

Neste contexto, é de fundamental importância que seja utilizado o mesmo algoritmo de optimização para os diferentes estimadores, de modo a assegurar que as diferenças de comportamento verificadas residam nas propriedades dos estimadores, e não nas características de um dado procedimento particular. Uma possibilidade é o algoritmo MDE, devido à sua flexibilidade na acomodação das funções objectivo associadas à definição dos diversos estimadores estudados, e, na nossa experiência, à elevada qualidade das soluções calculadas sem necessidade de regular os valores por omissão dos parâmetros de controlo.

Apesar das limitações do pequeno número de problemas estudados, o facto de no caso com uma só resposta ser possível identificar nitidamente um padrão partilhado por problemas de natureza bastante diferente, leva-nos a propor um quadro de conjecturas, em nosso entender, sólidas (veja-se para mais pormenores a secção 4.9 na página 102). Neste caso, a ideia geral subjacente aos métodos robustos é que constituem uma alternativa apropriada ao método dos mínimos quadrados, sendo competitivos mesmo em cenários Gaussianos, e confirmam, no essencial, as características apontadas no modelo linear. Os estudos de simulação mostram que não há um método que se possa dizer ser o melhor ou que exiba uma preponderância clara sobre os restantes em todas as situações.

Por outro lado, a análise do caso multivariado mostrou resultados mistos, o que dificulta a sua interpretação. Torna-se, portanto, necessário realizar estudos de caso adicionais para fixar ideias sobre os métodos robustos multivariados.

Para finalizar, as limitações existentes são essencialmente duas:

- Como se fez notar anteriormente, o pequeno número de experiências (estudos de caso) realizadas.
- O estudo aqui feito não analisa o efeito da dimensão da amostra. No esquema "clássico" para gerar uma amostra aleatória de tamanho n geram-se valores aleatórios para as variáveis independentes numa dada região. Ao contrário, o contexto do presente trabalho é o de experimentação em que é usado um procedimento formal ou informal de delineamento de experiências, problema que não será aqui tratado.

6.2 Direcções futuras

Nos parágrafos seguintes são referidos alguns tópicos de investigação futura.

Considerando o pequeno número de estudos de caso deste trabalho, fica claro que em primeiro lugar há necessidade de proceder a mais experimentação. Também é necessário cobrir outras categorias para além da cinética química e termodinâmica.

Um problema diz respeito à complexidade de muitos problemas reais. De facto, acima de certos limites o esforço computacional do método de Monte Carlo torna-se impraticável. Nestas circunstâncias, pensamos que as características paralelas inerentes quer ao método de Monte Carlo quer aos algoritmos de optimização do tipo computação evolutiva como o MDE, tornam particularmente prometedora a via da computação paralela para ultrapassar esta questão.

Outra hipótese consiste no estudo de estimadores e versões não incluidos neste trabalho, por exemplo: estimador S (Rousseeuw e Yohai, 1984; Sakata e White, 2001), estimador L_p aparado (Müller, 1997, p. 59), estimador WARME (Wang *et al.*, 2000). Uma interessante extensão dos estimadores aparados é a de Olive e Hawkins (2003) que descreve um método para escolha automática da proporção de aparamento.

Uma outra hipótese que se levanta, já referida, é a de avaliar a influência da dimensão amostral no desempenho dos estimadores.

Além disso, a situação de correlação entre respostas no caso multivariado descrita atrás na secção 2.4.3 na página 37 necessita ainda de ser devidamente resolvida.

Para terminar, chama-se a atenção para uma questão enfatizada em Gleser (1998). Segundo Gleser, os avanços de precisão conseguidos em diversos processos de medida pela inovação tecnológica, tornaram significativas fontes de erro sistemático (não elimináveis), no que se refere à sua magnitude quando confrontada com a magnitude do erro aleatório. Como orientação, Gleser refere o caso em que o desvio padrão do erro aleatório e o nível do erro sistemático são equivalentes. Repare-se que, no passado, era possível descartar o efeito deste tipo de erro sistemático.

Assim, Gleser advoga que o modelo de medição deve incluir um termo associado ao erro sistemático. A estrutura conceptual mais simples é

$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + \epsilon + b,$$

onde b representa o sinal e a magnitude do erro sistemático.

Ora, o estimador LTD revela-se particularmente adequado neste contexto, já que (secção 2.2.7 na página 24) é por inerência resistente ao erro sistemático. Por outro lado, o erro sistemático nas medições pode incorporar-se nos restantes estimadores mediante a especificação de b como um parâmetro (perturbador) adicional.

A desvantagem desta via em relação ao estimador LTD, é que exige o cálculo de uma estimativa do parâmetro perturbador b. Obviamente que isto não significa que a avaliação cabal do desempenho não seja obrigada a recorrer, mais uma vez, a um estudo de simulação de Monte Carlo análogo ao realizado neste trabalho.

Anexo A

Diagnóstico de outliers para o modelo de regressão linear

É sabido que os resíduos do estimador dos mínimos quadrados são deficientes na identificação de *outliers*. Por exemplo, a abordagem clássica, em que se comparam os resíduos padronizados pelo desvio padrão com um valor crítico mascara os *outliers*, pois o valor do desvio padrão estimado é inflacionado pela atracção que os *outliers* exercem sobre a curva de regressão. Outra forma do efeito de máscara que ocorre com o estimador dos mínimos quadrados é a camuflagem dos *outliers* quando estes se encontram agrupados. O uso dos resíduos de um ajuste robusto permite superar as deficiências acabadas de apontar. Assim, um subproduto do uso dum estimador robusto é poder ser empregue como ferramenta de diagnóstico.

Além da direcção da variável independente, as observações podem apresentar discordâncias do padrão da generalidade dos dados no espaço das variáveis independentes. Torna-se, por isso, igualmente necessário detectar os denominados *pontos de repercussão* (*leverage points*), isto é, pontos que são *outliers* em $\boldsymbol{x}$. Por outras palavras, $(y_i, \boldsymbol{x}_i)$ é um ponto de repercussão se $\boldsymbol{x}_i$ for uma observação discordante relativamente ao globo dos $\boldsymbol{x}_i$ nos dados. A detecção dos pontos de repercussão torna-se difícil quando a dimensão de $\boldsymbol{x}$ é superior a 2, pois deixa de ser possível recorrer à simples inspecção visual. O procedimento clássico consiste em procurar $\boldsymbol{x}_i$ com valores elevados da distância de Mahalanobis

$$\mathrm{MD}_i = \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{V}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})},$$

onde $\boldsymbol{\mu}$ é a média aritmética e $\boldsymbol{V}$ é a matriz de covariâncias amostral dos $\boldsymbol{x}_i$. No entanto, esta abordagem sofre igualmente dum efeito de máscara — múltiplos *outliers* não têm necessariamente um valor alto de MD_i —, uma vez que $\boldsymbol{\mu}$ e $\boldsymbol{V}$ não são robustos. Tal como sucede com o cálculo da variância dos resíduos resultantes do procedimento de regressão, um agrupamento de *outliers* atrai $\boldsymbol{\mu}$ e inflaciona $\boldsymbol{V}$ na sua direcção.

Posto isto, Rousseeuw e van Zomeren (1990) propuseram a representação gráfica dos resíduos dum estimador robusto com elevado ponto de rotura padronizados por uma estimativa robusta de escala[1] em função duma modificação robustificada da distância de Mahalanobis, RD_i, para a qual as estimativas de $\boldsymbol{\mu}$ e $\boldsymbol{V}$ são dadas por estimadores robustos de localização e covariância, tal como o estimador do MCD[2] (Rousseeuw e

[1] Este procedimento é particularmente adequado a estimadores que englobem uma estimativa robusta de escala na sua definição — como, por exemplo, os estimadores MM e τ descritos nas secções 2.2.5 na página 22 e 2.2.6 na página 23, respectivamente —, pois deste modo todas as estimativas necessárias são produto da regressão robusta.

[2] O uso deste estimador em detrimento do estimador do elipsóide de volume mínimo (MVE) usado por

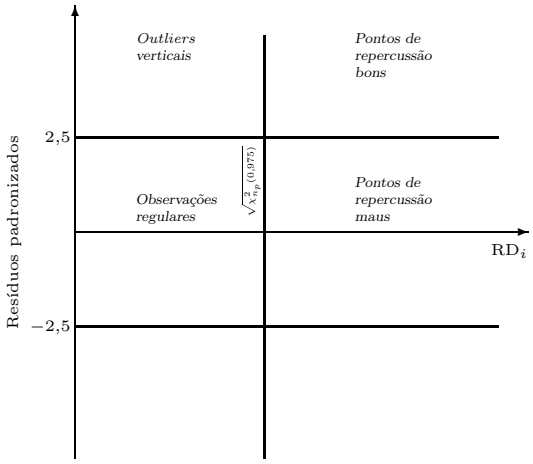


Figura A.1 Ilustração do gráfico de diagnóstico de Rousseeuw e van Zomeren.

Leroy, 1987, pp. 262–264). Os pontos situados à direita do valor de corte $\sqrt{\chi^2_{n_p}(0{,}975)}$ são considerados pontos de repercussão, enquanto aqueles que se encontram fora da banda de tolerância horizontal compreendida entre $[-2{,}5, 2{,}5]$ são *outliers* de regressão. O gráfico (ver figura A.1) permite distinguir quatro tipos de *outliers*:

Observações regulares, constituídas pelos pontos com distâncias robustas RD_i baixas (inferiores ao valor de corte) e resíduos padronizados baixos (no interior da banda de tolerância).

Outliers verticais, pontos com distâncias robustas RD_i baixas mas de resíduos padronizados elevados.

Pontos de repercussão bons, observações com distâncias RD_i elevadas e resíduos padronizados baixos. Estes pontos constituem *outliers* no espaço das variáveis independentes mas a variável dependente segue a curva de regressão.

Pontos de repercussão maus, observações com distâncias RD_i e resíduos padronizados elevados. Estes pontos têm em geral uma influência desastrosa na análise clássica do método dos mínimos quadrados, em que ficam mascarados.

(Para uma discussão de tipos de *outliers* veja-se igualmente a secção 11.2 de Ryan (1997).) Rousseeuw e van Zomeren (1990) chamam a atenção para o facto dos valores de corte serem em certa medida arbitrários, pelo que os casos situados na fronteira não apresentam uma classificação claramente definida. Os mesmos autores recomendam como regra prática o uso do estimador robusto de localização e covariância para $n/n_p > 5$ de modo a evitar o vazio do espaço multivariado (a chamada "maldição da dimensionalidade").

Uma extensão desta técnica de diagnóstico ao quadro de regressão em modelos com resposta multivariada é apresentada em Rousseeuw *et al.* (2000, p. 18). Neste caso,

Rousseeuw e van Zomeren (1990) é recomendado por Zaman *et al.* (2001), uma vez que o estimador MCD é mais vantajoso em termos de eficiência.

representa-se a distância robusta dada por $(\hat{\boldsymbol{\epsilon}}_i^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\epsilon}}_i)^{1/2}$ em função de RD_i; $\hat{\boldsymbol{\epsilon}}$ representa o vector dos resíduos de um estimador robusto, e $\hat{\boldsymbol{\Sigma}}$ é uma estimativa robusta da matriz de covariâncias dos erros de medição.

Anexo B

Descrição genérica das principais funções S3 utilizadas nas experiências de simulação

Com este anexo, pretende-se dar destaque a alguns dos programas mais relevantes utilizados nas experiências de simulação apresentadas no texto. O levantamento que fizemos sobre os argumentos das funções é selectivo: apenas incluímos os mais importantes.

B.1 Implementação computacional do algoritmo de optimização MDE

A chamada da função `mde` que se escreveu para implementar este algoritmo (tal como foi descrito na página 34) assume, tipicamente, a seguinte forma:

```
mde(lower, upper, f0, fn, , , NP, , , , tol, , maxiter, init, , , ...)
```

na qual

`lower`, `upper`	Vectores numéricos que introduzem limitações inferiores e superiores para os valores das `d` variáveis de decisão.
`f0`	Valor numérico correspondente à constante normalizadora f_0 usada na condição de paragem.
`fn`	Função objectivo a minimizar, em que o primeiro argumento é o vector de variáveis de decisão. Deve retornar uma quantidade escalar.
`NP`	Tamanho da população inicial gerada aleatoriamente. O valor por omissão é `10*d`.
`tol`	Tolerância para a condição de paragem. O valor por omissão é `1e-15`.
`maxiter`	Número máximo de gerações. O valor por omissão é `200*d`.
`init`	Vector numérico facultativo ou matriz com pontos candidatos *a juntar* à população inicial. Se fornecido como uma matriz, cada coluna especifica um ponto.
`...`	Argumentos facultativos passados à função `fn`.

A função retorna uma lista com os seguintes componentes:

`par` é o vector numérico com a melhor óptimo conseguido na geração final,

`value` é o valor da função objectivo correspondente a `par`,

`convergence` é um inteiro que retorna o valor `1` se o número máximo de gerações é atingido, caso contrário toma o valo `0`.

O uso da função `mde` vai ser ilustrado através de uma função de teste que se reproduz de Pínter (2002, p. 540)

$$f(\boldsymbol{x}) = 0{,}025d\sum_{i=1}^{d}(x_i - x_i^*)^2 + \sin^2\left[\sum_{i=1}^{d}(x_i - x_i^*) + \sum_{i=1}^{d}(x_i - x_i^*)^2\right] + \sin^2\left[\sum_{i=1}^{d}(x_i - x_i^*)\right],$$

onde d denota a dimensão do problema e $\boldsymbol{x}^*$ denota o óptimo global gerado aleatoriamente, em que cada elemento é escolhido a partir da distribuição uniforme em $[-5, 5]$. Como facilmente se verifica $f(\boldsymbol{x}^*) = 0$. Segue a descrição do código:

```
> source("programas.R")
> f <- function(x, a) {
+   x <- x - a
+   A <- sum(x)
+   B <- sum(x^2)
+   0.025*length(x)*B + sin(A + B)^2 + sin(A)^2
+ }
> set.seed(123)
> minimum <- runif(10, -5, 5)
> minimum
> mde(-5, 5, 1, f, tol = 1e-10, a = minimum)
```

B.2 Caso univariado

B.2.1 Funções para o modelo de regressão não-linear

Consideremos agora as funções usadas para o cálculo de estimativas de mínimos quadrados, ou robustas. Especificamente:

`nlrols` para ajuste pelo método dos mínimos quadrados,

`nlrlp` calcula as estimativas de norma L_p, quer quando p é conhecido quer quando p é desconhecido,

`nlrhbp` calcula as estimativas de um dos métodos LMS, LTS, τ ou LTD,

`nlrmm` calcula as estimativas MM.

Em cada caso, uma chamada típica toma a seguinte forma:

```
nlrols(, , , lower, upper, model, , data, , pnames, , , , , ...)
nlrlp(, interval, , , lower, upper, model, , data, , pnames, , , , ...)
nlrhbp(, method, trim, , , lower, upper, model, data, ,
```

```
       pnames, , , , ...)
nlrmm(, eff, , , , , , , , , lower, upper, model, , data, ,
      pnames, , , , ...)
```

onde

`lower`, `upper` Vectores numéricos que introduzem limitações inferiores e superiores para os valores das variáveis de decisão.

`model` Fórmula de especificação do modelo, que é definida do seguinte modo:

```
resposta ~ modelo
```

onde `modelo` é uma expressão aritmética que envolve os parâmetros e as variáveis independentes.

`data` *Data frame* que contém as variáveis especificadas em `model`, ou seja, os dados relativos a um problema.

`pnames` Vector carácter de nomes dos parâmetros do modelo.

`...` Argumentos facultativos passados à função de optimização.

`interval` Vector numérico de dois elementos para restringir o intervalo de variação admissível para p, ou escalar que permite fixar o valor actual de p. O valor por omissão é `c(1, 10)`.

`method` Variável carácter que especifica o método de ajuste. Os valores possíveis são `"lms"` para o estimador LMS, `"lts"` para o estimador LTS, `"tau"` para o estimador τ, e `"ltd"` para o estimador LTD. O valor por omissão é `"lts"`.

`trim` Proporção de aparamento, quando `method = "ltd"`. O valor por omissão é `0.25`.

`eff` Eficiência (em percentagem) sob erro de medição Gaussiano. O valor por omissão é `95`.

Para efeitos de ilustração, vamos considerar o processo de geração de dados usado na simulação de Monte Carlo feita por Midi (1999, pp. 596 e 597). O modelo que vamos escolher é o seguinte:

$$y_i = 6\exp[-\exp(0{,}7 - 0{,}4x_i)] + \epsilon_i,$$

onde x_i é gerado a partir da distribuição uniforme no intervalo $[1, 10]$ e ϵ_i é N(0, 0,1).

Partamos do seguinte conjunto de dados constituído por 25 observações:

```
> source("programas.R")
> library(combinat)
> gompertz <- function(b0, b1, b2, x) b0*exp( -exp(b1 - b2*x) )
> sigma <- sqrt(0.1)
> set.seed(123)
```

```
> x1 <- runif(25, 1, 10)
> err <- rnorm(length(x1), sd = sigma)
> Gomp <- data.frame( x = x1, y = gompertz(6, 0.7, 0.4, x1) + err )
> fm1Gomp.ols <- nlrols( lower = 0, upper = 20,
+                        model = y ~ gompertz(b0, b1, b2, x),
+                        data = Gomp, pnames = c("b0", "b1", "b2") )
> coef(fm1Gomp.ols)
> fm1Gomp.tau <- nlrhbp( method = "tau", lower = 0, upper = 20,
+                        model = y ~ gompertz(b0, b1, b2, x),
+                        data = Gomp, pnames = c("b0", "b1", "b2") )
> coef(fm1Gomp.tau)
```

Em seguida, consideremos os mesmos dados mas em vez de usarmos a primeira observação anterior, vamos agora substituí-la por um *outlier* com x_1 gerado a partir da distribuição uniforme em $[1, 2]$ e y_1 é N(10, 0,1).

```
> bad <- 1
> Gomp$x[bad] <- runif(length(bad), 1, 2)
> Gomp$y[bad] <-
+     gompertz(6, 0.7, 0.4, dat$x[bad]) + rnorm(length(bad), 10, sigma)
> coef(update(fm1Gomp.ols))
> coef(update(fm1Gomp.tau))
```

B.2.2 Simulação de Monte Carlo

A função `monte.carlo.robreg.simulations` que executa as experiências de Monte Carlo descritas na secção 3.2 na página 40 pode ser usada da seguinte maneira:

```
monte.carlo.robreg.simulations(object, nsim = 500,
    vertical.outliers = FALSE, contamination.fraction = 0.10,
    factor.y = 5, ...)
```

onde

`object` Um objecto correspondente a um ajuste pelo método dos mínimos quadrados.

`nsim` Número de réplicas.

`vertical.outliers`
Valor lógico que permite escolher se a amostra é contaminada ou não.

`contamination.fraction`
Proporção de contaminação da amostra.

`factor.y` Número de desvios padrão do erro que os *outliers* distam dos pontos regulares no espaço da variável dependente.

`...` Argumentos passados às funções para ajuste de parâmetros.

O procedimento de cálculo dos intervalos de confiança dos índices de desempenho é implementado na função `simul.ci`. A chamada é

```
simul.ci(x, nboot = 999, conf = 0.95, type = "perc")
```

na qual

`x`	Um objecto correspondente a uma simulação de Monte Carlo.
`nboot`	Número de réplicas *bootstrap*.
`conf`	Grau ou coeficiente de confiança.
`type`	Variável carácter que especifica o procedimento de cálculo do intervalo de confiança. O valor por omissão selecciona o método do percentil. Para detalhes veja-se Davison e Hinkley (1997, cap. 5).

Por exemplo, os comandos:

```
> library(boot)
> library(sn)
> fm1Gomp.sim <- monte.carlo.robreg.simulations( fm1Gomp.ols,
+     nsim = 100, vertical.outliers = TRUE,
+     contamination.fraction = 0.15, factor.y = 10,
+     lower = 0, upper = 20, pnames = c("b0", "b1", "b2") )
> fm1Gomp.sim.ci <- simul.ci(fm1Gomp.sim)
```

executam um estudo de Monte Carlo para avaliar o desempenho dos estimadores na presença de *outliers* nos dados.

A representação gráfica dos resultados pode ser obtida usando as funções `plot.robsim` e `plot.robsimci`.

```
> plot(fm1Gomp.sim)
> plot(fm1Gomp.sim.ci)
```

B.3 Caso multivariado

B.3.1 Funções para o modelo de regressão não-linear

Agora, com duas ou mais variáveis dependentes, estão disponíveis as funções:

`mvnlr`, que calcula as estimativas referentes ao critério do determinante,

`mvnlrob`, que calcula as estimativas de um dos métodos LAD multivariado, M multivariado, MML ou MTL.

A utilização é feita do seguinte modo:

```
mvnlr(, , , lower, upper, model, , data, , pnames, , , , ...)
mvnlrob(, method, k, trim, , , , , , , , , , , , , , lower, upper,
        lower.sigmaii, upper.sigmaii, model, data, , pnames, , , ...)
```

onde

`lower`, `upper` Vectores numéricos que introduzem limitações inferiores e superiores para os valores das variáveis de decisão.

`model` Fórmula de especificação do modelo, que é definida do seguinte modo:

```
c(resp1,...,respn) ~ modelo
```

onde `resp1`,...,`respn` representa as variáveis dependentes e `modelo` é uma expressão aritmética que envolve os parâmetros e as variáveis independentes. Este último deve retornar uma matriz reunindo todos os vectores coluna dos valores ajustados para as variáveis dependentes, ordenados como em `resp1`,...,`respn`.

`data` *Data frame* que contém as variáveis especificadas em `model`, ou seja, os dados relativos a um problema.

`pnames` Vector carácter de nomes dos parâmetros do modelo.

`...` Argumentos facultativos passados à função de optimização.

`method` Variável carácter que especifica o método de ajuste. Os valores possíveis são `"lad"` para o estimador LAD multivariado, `"m"` para o estimador M multivariado, `"mtl"` para o estimador MTL, e `"mml"` para o estimador MML. O valor por omissão é `"lad"`.

`k` Nó da função de Huber, quando `method = "m"`. O valor por omissão é `1.5`.

`trim` Proporção de aparamento, quando `method = "mtl"`. O valor por omissão é `0.25`.

`lower.sigmaii`, `upper.sigmaii`
Vectores numéricos que introduzem limitações inferiores e superiores para os valores das variâncias do erro de medição. A gama de variação é, por omissão, `1e-10`,... e `mad(resp1, 0)`,...,`mad(respn, 0)`.

A título de ilustração, consideremos um exemplo retirado de Esposito e Floudas (1998, exemplo 4, pp. 1850–1852). O modelo é especificado do seguinte modo:

$$y_i = \left(\theta_1 + \frac{\theta_2}{x_i^{\theta_3}}\right) + \mathrm{j}\left(x_i\theta_4 - \frac{\theta_5}{x_i^{\theta_3}}\right),$$

onde j é o número imaginário $\mathrm{j} = \sqrt{-1}$ e $x_i = i\pi/20$. Considere-se o seguinte conjunto de dados (Esposito e Floudas, 1998, tabela 19, p. 1856):

i	1	2	3	4	5	6
y_i	$5-5\mathrm{j}$	$3-2\mathrm{j}$	$2-\mathrm{j}$	$1{,}5-0{,}5\mathrm{j}$	$1{,}2-0{,}2\mathrm{j}$	$1{,}1-0{,}1\mathrm{j}$

```
> source("programas.R")
> library(nlme)
> y <- c(5 - 5i, 3 - 2i, 2 - i, 1.5 - 0.5i, 1.2 - 0.2i, 1.1 - 0.1i)
> Resp <- data.frame( x = (1:6)*pi/20, yreal = Re(y), yimag = Im(y) )
> lo <- c(0, 0, 1.1, 0, 0)
> hi <- c(1, 1, 1.3, 1, 1)
> fm1Resp.det <- mvnlr( lower = lo, upper = hi,
+     model = c(yreal, yimag) ~ cbind(b1 + b2/x^b3, x*b4 - b5/x^b3),
+     data = Resp, pnames = c("b1", "b2", "b3", "b4", "b5") )
> coef(fm1Resp.det)
> fm1Resp.lad <- mvnlrob( method = "lad", lower = lo, upper = hi,
+     model = c(yreal, yimag) ~ cbind(b1 + b2/x^b3, x*b4 - b5/x^b3),
+     data = Resp, pnames = c("b1", "b2", "b3", "b4", "b5") )
> coef(fm1Resp.lad)
```

B.3.2 Simulação de Monte Carlo

O uso das funções `monte.carlo.mvrobreg.simulations`, `mv.simul.ci`, `plot.mvrobsim`, e `plot.mvrobsimci`, desenvolvidas para o caso multivariado é, *mutatis mutandis*, semelhante ao das funções relacionadas do caso univariado.

```
> library(boot)
> library(mvtnorm)
> fm1Resp.sim <- monte.carlo.mvrobreg.simulations( fm1Resp.det,
+     nsim = 100, vertical.outliers = TRUE,
+     contamination.fraction = 0.15, factor.y = 10,
+     lower = lo, upper = hi, pnames = c("b1", "b2", "b3", "b4", "b5") )
> fm1Resp.sim.ci <- mv.simul.ci(fm1Resp.sim)
> plot(fm1Resp.sim)
> plot(fm1Resp.sim.ci)
```

Bibliografia

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171–178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica* **46**, 199–208.

Bates, D. M. e Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York.

Beaton, A. E. e Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on hand-spectroscopic data (with discussion). *Technometrics* **16**, 147–192.

Becker, R. A., Chambers, J. M., e Wilks, A. R. (1988). *The NEW S Language*. Chapman & Hall, New York. (Publicado originariamente por Wadsworth and Brooks/Cole, Pacific Grove, CA.).

Belohlav, Z., Zamostny, P., Kluson, P., e Volf, J. (1997). Application of random-search algorithm for regression analysis of catalytic hydrogenations. *The Canadian Journal of Chemical Engineering* **75**, 735–742.

Bobbo, S., Camporese, R., e Stryjek, R. (1998). (Vapour + liquid) equilibrium measurements and correlations of the refrigerant mixture {dimethyl ether (RE170) + 1,1,1,3,3,3-hexafluoropropane (R236fa)} at the temperatures (303.68 and 323.75) K. *Journal of Chemical Thermodynamics* **30**, 1041–1046.

Box, G. E. P. e Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Massachusetts.

Brent, R. P. (1973). *Algorithms for Minimization Without Derivatives*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey. (Publicado em 2002 com o mesmo título, por Dover Publications, Mineola, New York.).

Byrd, R. H., Lu, P., Nocedal, J., e Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16**, 1190–1208.

Carr, N. L. (1960). Kinetics of catalytic isomerization of n-pentane. *Industrial and Engineering Chemistry* **52**, 391–396.

Cela, R., Martinéz, E., e Carro, A. M. (2001). Supersaturated experimental designs: New approaches to building and using it. Part II. Solving supersatured designs by genetic algorithms. *Chemometrics and Intelligent Laboratory Systems* **57**, 75–92.

Chambers, J. M. (1998). *Programming with Data. A Guide to the S Language*. Springer-Verlag, New York.

Chambers, J. M. e Hastie, T. J. (eds.) (1992). *Statistical Models in S*. Chapman & Hall, New York. (Publicado originariamente por Wadsworth and Brooks/Cole, Pacific Grove, CA.).

Chen, Y.-L., Stromberg, A. J., e Zhou, M. (1997). The least trimmed squares estimate in nonlinear regression. Rel. téc., Department of Statistics, University of Kentucky, Lexington, KY 40506.

Clancy, V. J. (1947). Statistical methods in chemical analyses. *Nature* **159**, 339–340.

Coello, C. A. C. (2002). Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: A survey of the state of the art. *Computer Methods in Applied Mechanics and Engineering* **191**, 1245–1287.

Croux, C., Rousseeuw, P. J., e Hössjer, O. (1994). Generalized S-estimators. *Journal of the American Statistical Association* **89**, 1271–1281.

Davison, A. C. e Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

De Santis, R., Gironi, F., e Marrelli, L. (1976). Vapor-liquid equilibrium from a hard-sphere equation of state. *Industrial and Engineering Chemistry Fundamentals* **15**, 183–189.

Donoho, D. L. e Huber, P. J. (1983). The notion of breakdown point. In P. J. Bickel, K. A. Doksum, e J. L. Hodges, Jr (eds.) *A Festschrift for Erich L. Lehmann*. Wadsworth, Belmont, CA, págs. 157–184.

Dutter, R. e Huber, P. J. (1981). Numerical methods for the nonlinear robust regression problem. *Journal of Statistical Computation and Simulation* **13**, 79–113.

Edgar, T. F., Himmelblau, D. M., e Lasdon, L. S. (2001). *Optimization of Chemical Processes*. 2ª ed. McGraw-Hill, New York.

Edlund, O., Ekblom, H., e Madsen, K. (1997). Algorithms for non-linear M-estimation. *Computational Statistics* **12**, 373–383.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Efron, B. e Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Nº 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton.

Englezos, P. e Kalogerakis, N. (2001). *Applied Parameter Estimation for Chemical Engineers*. Nº 81 in Chemical Industries. Marcel Dekker, New York.

Esposito, W. R. e Floudas, C. A. (1998). Global optimization in parameter estimation of nonlinear algebraic models via the error-in-variables approach. *Industrial and Engineering Chemistry Research* **37**, 1841–1858.

Field, C. e Ronchetti, E. (1990). *Small Sample Asymptotics*. Monograph Series. IMS Lecture Notes, Hayward, CA.

Floudas, C. A., Pardalos, P. M., Adjiman, C. S., Esposito, W. R., Gümüş, Z. H., Harding, S. T., Klepeis, J. L., Meyer, C. A., e Schweiger, C. A. (1999). *Handbook of Test Problems in Local and Global Optimization*. Nº 33 in Nonconvex Optimization and Its Applications. Kluwer Academic Publishers, Dordrecht.

Frame, K. K. e Hu, W.-S. (1988). A model for density-dependent growth of anchorage-dependent mammalian cells. *Biotechnology and Bioengineering* **32**, 1061–1066.

Gentle, J. E. (1998). *Random Number Generation and Monte Carlo Methods*. Statistics and Computing. Springer-Verlag, New York.

Gleser, L. J. (1998). Assessing uncertainty in measurement. *Statistical Science* **13**, 277–290.

Gonin, R. e Money, A. H. (1989). *Nonlinear L_p-Norm Estimation*, volume 100 de *Statistics: Textbooks and Monographs*. Marcel Dekker, New York.

Hadi, A. S. e Luceño, A. (1997). Maximum trimmed likelihood estimators: A unified approach, examples, and algorithms. *Computational Statistics and Data Analysis* **25**, 251–272.

Hampel, F. (1998). Is statistics too difficult? *The Canadian Journal of Statistics* **26**, 497–513.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393.

Hawboldt, K. A., Kalogerakis, N., e Behie, L. A. (1994). A cellular automaton model for microcarrier cultures. *Biotechnology and Bioengineering* **43**, 90–100.

Hawkins, D. M. (1994). The feasible solution algorithm for least trimmed squares regression. *Computational Statistics and Data Analysis* **17**, 185–196.

Hindmarsh, A. C. (1983). ODEPACK, a systematized collection of ODE solvers. In R. S. Stepleman (ed.) *Scientific Computing*. North-Holland, Amsterdam, págs. 55–64.

Hubbard, A. B. e Robinson, W. E. (1950). A thermal decomposition study of Colorado oil shale. Report of Investigation 4744, U. S. Bureau of Mines.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73–101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics* **1**, 799–821.

Huber, P. J. (1996). *Robust Statistical Procedures*. Nº 68 in CBMS-NSF Regional Conference Series in Applied Mathematics, 2ª ed. SIAM, Philadelphia.

Huet, S., Bouvier, A., Gruet, M.-A., e Jolivet, E. (1996). *Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS Examples*. Springer-Verlag, New York.

Ihaka, R. e Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314.

Johnston, J. e DiNardo, J. (2001). *Métodos Econométricos*. McGraw-Hill, Lisboa. (Tradução de Econometric Methods, 4.ª edição, McGraw-Hill, NY, 1997, por Manuela Magalhães Hill e outros.).

Jovanović, S. e Paunović, R. (1984). Generating appropriate density values from a cubic equation of state to avoid false unit K values. Application to distillation problems. *Industrial and Engineering Chemistry Process Design and Development* **23**, 801–805.

Kaufman, E. H., Jr, Taylor, G. D., Mielke, P. W., Jr, e Berry, K. J. (2002). An algorithm and Fortran program for multivariate LAD (ℓ_1 of ℓ_2) regression. *Computing* **68**, 275–287.

Kelley, C. T. (1999). *Iterative Methods for Optimization*. SIAM, Philadelphia.

Koenker, R. e Park, B. J. (1992). An interior point algorithm for nonlinear quantile regression. Rel. Téc. Faculty Working Paper, No. 92-0127, College of Commerce and Business Administration, University of Illinois at Urbana-Champaign.

Koenker, R. e Portnoy, S. (1990). M-estimation of multivariate regressions. *Journal of the American Statistical Association* **85**, 1060–1068.

Kolda, T. G., Lewis, R. M., e Torczon, V. (2003). Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review* **45**, 385–482.

Krishnakumar, J. e Ronchetti, E. (1997). Robust estimators for simultaneous equations models. *Journal of Econometrics* **78**, 295–314.

Křivý, I. e Tvrdík, J. (1995). The controlled random search algorithm in optimizing regression models. *Computational Statistics and Data Analysis* **20**, 229–234.

Křivý, I. e Tvrdík, J. (1999). Simple evolutionary heuristcs for global optimization. *Computational Statistics and Data Analysis* **30**, 345–352.

Křivý, I., Tvrdík, J., e Krpec, R. (2000). Stochastic algorithms in nonlinear regression. *Computational Statistics and Data Analysis* **33**, 277–290.

Lawrence, K. D. e Arthur, J. L. (1990). Robust nonlinear regression. In K. D. Lawrence e J. L. Arthur (eds.) *Robust Regression: Analysis and Applications*. Marcel Dekker, New York, págs. 59–86.

Le Cam, L. M. e Yang, G. L. (2000). *Asymptotics in Statistcs: Some Basic Concepts*. 2ª ed. Springer-Verlag, New York.

Lee, M. H., Han, C., e Chang, K. S. (1999). Dynamic optimization of a continuous polymer reactor using a modified differential evolution algorithm. *Industrial and Engineering Chemistry Research* **38**, 4825–4831.

Lee, M.-J., Wu, F.-L., e mu Lin, H. (2001). Separation of closely boiling compounds of catechol and 4-methoxyphenol with the aid of 1,4-butanediol. *Industrial and Engineering Chemistry Research* **40**, 4596–4602.

Lewis, P. A. W. e Orav, E. J. (1989). *Simulation Methodology for Statisticians, Operations Analysts, and Engineers*, volume I. Wadsworth and Brooks/Cole, Pacific Grove, CA.

Lohmann, T., Bock, H. G., e Schlŏder, J. P. (1992). Numerical methods for parameter estimation and optimal experiment design in chemical reaction systems. *Industrial and Engineering Chemistry Research* **31**, 54–57.

Maria, G. (1989). An adaptive strategy for solving kinetic model concomitant estimation-reduction problems. *The Canadian Journal of Chemical Engineering* **67**, 825–832.

Midi, H. (1999). Preliminary estimators for robust non-linear regression estimation. *Journal of Applied Statistics* **26**, 591–600.

Militký, J. e Čáp, J. (1987). Application of the Bayes approach to adaptive L_p nonlinear regression. *Computational Statistics and Data Analysis* **5**, 381–389.

Müller, C. H. (1997). *Robust Planning and Analysis of Experiments*, volume 124 de *Lecture Notes in Statistics*. Springer-Verlag, New York.

Murteira, B., Ribeiro, C. S., Silva, J. A., e Pimenta, C. (2001). *Introdução à Estatística*. McGraw-Hill, Lisboa.

Neugebauer, S. P. (1996). *Robust Analysis of M-Estimators of Nonlinear Models*. Tese de Mestrado, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Olive, D. J. e Hawkins, D. M. (2003). Robust regression with high coverage. *Statistics and Probability Letters* **63**, 259–266.

Phillips, G. R. e Eyring, E. M. (1983). Comparison of conventional and robust regression in analysis of chemical data. *Analytical Chemistry* **55**, 1134–1138.

Pínter, J. D. (2002). Global optimization: Software, test problems, and applications. In P. M. Pardalos e H. E. Romeijn (eds.) *Handbook of Global Optimization, Volume 2*, nº 62 in Nonconvex Optimization and Its Applications. Kluwer Academic Publishers, Dordrecht, págs. 515–569.

Price, K. V. (1999). An introduction to differential evolution. In D. Corne, M. Dorigo, e F. Glover (eds.) *New Ideias in Optimization*, Advanced Topics in Computer Science. McGraw-Hill, London, págs. 79–108.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79**, 871–880.

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

Rousseeuw, P. J., van Aelst, S., e van Driessen, K. (2000). Robust multivariate regression. Rel. téc., Department of Mathematics and Computer Science, University of Antwerp, Universiteitsplein 1, B-2610 Wilrijk, Belgium.

Rousseeuw, P. J. e van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points (with comments and rejoinder). *Journal of the American Statistical Association* **85**, 633–651.

Rousseeuw, P. J. e Yohai, V. (1984). Robust regression by means of *S*-estimators. In J. Franke, W. Härdle, e R. D. Martin (eds.) *Robust and Nonlinear Time Series Analysis*, nº 26 in Lecture Notes in Statistics. Springer-Verlag, New York, págs. 256–272.

Ryan, T. P. (1997). *Modern Regression Methods*. Wiley, New York.

Sakata, S. e White, H. (1995). An alternative definition of finite-sample breakdown point with applications to regression model estimators. *Journal of the American Statistical Association* **90**, 1099–1106.

Sakata, S. e White, H. (2001). *S*-estimation of nonlinear regression models with dependent and heterogeneous observations. *Journal of Econometrics* **103**, 5–72.

Sandler, S. I. (1999). *Chemical and Engineering Thermodynamics*. 3ª ed. Wiley, New York.

Schnabel, R. B., Koontz, J. E., e Weiss, B. E. (1985). A modular system of algorithms for unconstrained minimization. *ACM Transactions on Mathematical Software* **11**, 419–440.

Seber, G. A. F. e Wild, C. J. (1989). *Nonlinear Regression*. Wiley, New York.

Sebert, D. M., Montgomery, D. C., e Rollier, D. A. (1998). A clustering algorithm for identifying multiple outliers in linear regression. *Computational Statistics and Data Analysis* **27**, 461–484.

Staudte, R. G. e Sheather, S. J. (1990). *Robust Estimation and Testing*. Wiley, New York.

Storn, R. e Price, K. (1997). Differential evolution — a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization* **11**, 341–359.

Stromberg, A. J. (1993). Computation of high breakdown nonlinear regression parameters. *Journal of the American Statistical Association* **88**, 237–244.

Stromberg, A. J. (1995). Consistency of the last median of squares estimator in nonlinear regression. *Communications in Statistics—Theory and Methods* **24**, 1971–1984.

Stromberg, A. J. (1997a). Robust covariance estimates based on resampling. *Journal of Statistical Planning and Inference* **57**, 321–334.

Stromberg, A. J. (1997b). Some software for computing robust linear or nonlinear regression estimators. *Communications in Statistics—Simulation and Computation* **26**, 947–959.

Stromberg, A. J., Hössjer, O., e Hawkins, D. M. (2000). The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association* **95**, 853–864.

Stromberg, A. J. e Ruppert, D. (1992). Breakdown in nonlinear regression. *Journal of the American Statistical Association* **87**, 991–997.

Tabatabai, M. A. e Argyros, I. K. (1993). Robust estimation and testing for general nonlinear regression models. *Applied Mathematics and Computation* **58**, 85–101.

Tan, H., Su, X., Wei, W., e Yao, S. (1999). Robust complex non-linear regression method for the estimation of equivalent circuit parameters of the thickness-shear-mode acoustic wave sensor. *Chemometrics and Intelligent Laboratory Systems* **48**, 71–80.

Tan, H. S., Downie, J., e Bacon, D. W. (1988). The kinetics of the oxidation of propylene over a bismuth molybdate catalyst. *The Canadian Journal of Chemical Engineering* **66**, 611–618.

Vegliò, F., Trifoni, M., Pagnanelli, F., e Toro, L. (2001a). Shrinking core model with variable activation energy: A kinetic model of manganiferous ore leaching with sulphuric acid and lactose. *Hydrometallurgy* **60**, 167–179.

Vegliò, F., Trifoni, M., e Toro, L. (2001b). Leaching of manganiferous ores by glucose in a sulfuric acid solution: Kinetic modeling and related statistical analysis. *Industrial and Engineering Chemistry Research* **40**, 3895–3901.

Venables, W. N. e Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS*. Statistics and Computing, 3ª ed. Springer-Verlag, New York.

Wang, D., Romagnoli, J. A., e Safavi, A. A. (2000). Wavelet-based adaptive robust M-estimator for nonlinear system identification. *AIChE Journal* **46**, 1607–1615.

Watts, D. G. (1994). Estimating parameters in nonlinear rate equations. *The Canadian Journal of Chemical Engineering* **72**, 701–710.

Wehrens, R., Putter, H., e Buydens, L. M. C. (2000). The bootstrap: a tutorial. *Chemometrics and Intelligent Laboratory Systems* **54**, 35–52.

Wisnowski, J. W., Montgomery, D. C., e Simpson, J. R. (2001). A comparative analysis of multiple outlier detection procedures in the linear regression model. *Computational Statistics and Data Analysis* **36**, 351–382.

Wu, X. e Çinar, A. (1996). An adaptive robust M-estimator for nonparametric nonlinear system identification. *Journal of Process Control* **6**, 233–239.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics* **15**, 642–656.

Yohai, V. J. e Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association* **83**, 406–413.

You, J. (1999). A Monte Carlo comparison of several high breakdown and efficient estimators. *Computational Statistics and Data Analysis* **30**, 205–219.

Zaman, A., Rousseeuw, P. J., e Orhan, M. (2001). Econometric applications of high-breakdown robust regression techniques. *Economics Letters* **71**, 1–8.

Zhu, X. D., Valerius, G., Hofmann, H., Haas, T., e Arntz, D. (1997). Intrinsic kinetics of 3-hydroxypropanal hydrogenation over $Ni/SiO_2/Al_2O_3$ catalyst. *Industrial and Engineering Chemistry Research* **36**, 2897–2902.

Ziegel, E. R. e Gorman, J. W. (1980). Kinetic modelling with multiresponse data. *Technometrics* **22**, 139–151.