AUSTRIAN JOURNAL OF STATISTICS
Volume 31 (2002), Number 2&3, 131-140

Bayesian Inference for Questionable Data

Klaus Felsenstein
Department of Statistics and Probability Theory
Vienna University of Technology

Abstract: In this paper we develop Bayesian procedures for vague data. The data are assumed to be vague in the sense that the likelihood is a mixture of the model distribution and an error distribution. In this case the standard updating procedure of the model prior would fail.

As a new method to deal with such imprecise data we consider *observable* uncertainties. In this model a specified degree of belief for the validity of the observation is added to the original measurement. Our proposal involves the idea that occasionally the observations are caused by an unknown error distribution. We discuss the effect of this assumption and show a parametrical and non-parametrical analysis in this setup.

For the analysis of the error distribution we establish a nonparametrical approach. Convex optimization procedures can be applied for a nonparametric estimation of the error distribution. An equivalence theorem characterizes optimal estimates and provides an iterative procedure converging to the empirical Bayes estimate.

Keywords: Empirical Bayes Estimate, Mixture Models, Convex Optimization.

1 Introduction

In many practical applications the correctness of the data has to be checked. Assume a model with density $f(x|\theta)$ represents the standard case but the data are contaminated. In some cases data x are reported that are not informative about θ and the standard model. The basic concept of robust methods consists of replacing the original distribution $f(x|\theta)$ by a mixture model

$$f(x) \;=\; \alpha f(x|\theta) \;+\; (1-\alpha)g(x) \tag{1}$$

with a distribution of distortion $g(x)$ and a mixture weight $0 \leq \alpha \leq 1$.

Such mixture models are the starting point for robust methods. From a Bayesian point of view the robustness may be considered in various ways (Berger, 1994; Berger et al., 1996). The unknown mixture weight can be treated as a constant or a stochastic parameter with a prior distribution. Frequently, Bayesian robustness is understood as insensitivity to disturbed priors (Berger and Berliner, 1986; Bayarri and Berger, 1998; Ruggeri and Wasserman, 1993). An alternative approach of Bayesian robustness represents the analysis of changes in the loss function (DeGroot, 1988). Anyway, a mixture model gives rise to several difficulties concerning the estimation of the weight as well as determination of the distribution $g(x)$ (Lavine et al., 1993).

From a Bayesian point of view it seems natural to consider the mixture weight to be a stochastic variable (Gustafson and Wasserman, 1995). In our setup this stochastic variable is observable in the following sense.

As a delimitation to other robust methods let's give a definition of *questionable data*. The data x are reported by an expert who is able to evaluate the authenticity of x. We mean the data x is questionable if we are able to ask the expert how much he relies on the correctness of the data. The expert reports a probability b for each x. b is the probability that x is distributed according to the (assumed) model $f(x|\theta)$. If $b = 1$ the expert is sure that x is a correct measurement and if $b = 0$ the expert is sure that x has nothing to do with the assumed model. Apart from that b will be between 0 and 1.

Therefore we add a value of belief b_i with $0 \leq b_i \leq 1$ to each observation x_i. b_i (hopefully near 1) represents a realization of the stochastic variable α. Questionable data are pairs of measurements and beliefs, (x_i, b_i). Note that there exist completely different (non-probabilistical) methods for analysing imprecise measurements such as belief functions or "fuzzy data".

Since we are interested in the parameter θ of the original model we discuss the calculation of the posterior distribution of θ. How are the beliefs used as weights of the measurements x_i properly. First we consider the algorithm of updating especially the calculation of the posterior and the Bayes estimate of θ.

2 Updating and Estimation

The entire data frame $D = (x, b)$ consists of n observations x_i, $i = 1, \ldots, n$, and associated beliefs $b = (b_1, \ldots, b_n)$. First we consider a fixed distribution $g(x)$. Let $h(.)$ denote the density of α. We assume that the prior distribution $\pi(\theta)$ is independent of α and $g(x)$.

Let $m(x)$ denote the marginal density of the observation x under the original model $f(x|\theta)$, $m(x) = \int f(x|\theta)\pi(\theta)\, d\theta$. If no uncertainty is present, the prior $\pi(\theta)$ is updated to $\pi(\theta|x)$ following Bayes theorem,

$$\pi(\theta|D) \;=\; \pi(\theta|x) \;=\; \frac{f(x|\theta)\pi(\theta)}{m(x)}.$$

For $n = 1$ questionable observation $D = (x, b)$, the density $f(x|\theta)$ is replaced by the mixture density, therefore the posterior density reads

$$\pi(\theta|D) \;=\; \frac{\pi(\theta|x)m(x)b \;+\; (1-b)g(x)\pi(\theta)}{m(x)b \;+\; (1-b)g(x)}. \tag{2}$$

Since the posterior density for questionable observation is a weighted average of the posterior density with exact data, $\pi(\theta|x)$, and the prior $\pi(\theta)$ the Bayes estimate of θ results as the analogue average

$$\hat{\theta} \;=\; \frac{\theta_B(x)m(x)b \;+\; (1-b)g(x)\mu_0}{m(x)b \;+\; (1-b)g(x)} \tag{3}$$

with prior mean $\mu_0 = \mathbb{E}(\theta)$ and posterior mean $\theta_B(x) = \mathbb{E}(\theta|x)$ respectively. The updating for n questionable observations D follows

$$\pi(\theta|D) \;=\; \sum \omega(i_1, \ldots, i_k)\pi(\theta|x_{i_1}, \ldots, x_{i_k}) \tag{4}$$

where the sum is taken over all subsets of indices $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ and the weights ω are up to a normalizing constant

$$\omega(i_1, \ldots, i_k) \propto b_{i_1} ... b_{i_k} m(x_{i_1}, \ldots, x_{i_k}) \times (1 - b_{i_{k+1}}) ... (1 - b_{i_n}) g(x_{i_{k+1}}) ... g(x_{i_n}).$$

The Bayes estimate is the corresponding convex combination

$$\hat{\theta} = \sum \omega(i_1, \ldots, i_k) \theta_B(x_{i_1}, \ldots, x_{i_k}).$$

In the situation of conjugate priors the calculation of a marginal density reduces to the calculation of a quotient of prior and posterior densities. If n gets large a stepwise updating procedure according to (2) is recommended.

In practical situations assuming a fixed distortion distribution $g(.)$ becomes unrealistic. First, we consider a parametric family for $g(.)$. If a parameter τ determines the density $g(x|\tau)$ the specification of a additional prior distribution $\pi(\tau)$ is needed. Then the above updating procedures remain valid if we insert the marginal density

$$g(x) = \int g(x|\tau) \; d\pi(\tau)$$

with respect to the prior $\pi(\tau)$. The next level of difficulty is to incorporate an appropriate prior $\pi(\tau)$ for the nuisance parameter τ. Typically there is rare knowledge about the distortion distribution. Frequently the usage of a non-informative prior can not solve that problem. The non-informative (invariance) property depends upon the assessed family of distributions.

Instead we refuse to specify a special prior $\pi(\tau)$ and turn to a nonparametric approach concerning the distortion distribution. We consider a mixture model of the form

$$f(x) = \alpha f(x|\theta) + (1 - \alpha) \tilde{g}(x)$$

with unknown measure $\tilde{G}$. $\tilde{g}$ denotes the density of $\tilde{G}$. In a linear representation we write the measure as

$$f(x) = \alpha f(x|\theta) + (1 - \alpha) \int g_s(x) \, dG(s) \tag{5}$$

with a family of distributions $g_s(.)$. The model (5) serves as starting point for the estimation of the measure $G(.)$, it represents a completely unknown prior. Note that the distribution $h(.)$ of α does not affect the estimation of θ explicitely.

3 Estimation of the Distortion

There are several approaches towards the estimation of the measure G in (5). Instead of assessing a prior over measures (Antoniak, 1974) we introduce an empirical Bayes procedure here. The basic conception consists of a likelihood estimation of G first and a Bayesian estimation of θ according the guide line of Section 2 afterwards.

Before we discuss the Bayesian estimation we sketch the strategy of maximizing a nonparametric likelihood. Let the parameter of interest θ be fixed. The observation of questionable data leads to the likelihood

$$L(G) = \prod_i f(x_i|b_i)$$

and $l(G) = \log(L)$ the Log-likelihood. For the optimization of l standard procedures of convex optimization are used (Lindsay, 1995). The concave function $l()$ (in G) can be maximized over a convex set.

Define the feasible region $\mathcal{P}^*$ to be the convex hull of

$$\mathcal{P} \;=\; \{f(x_i|b_i); i = 1, \ldots, n, G\}.$$

The feasible region is a subset of $\mathbb{R}^n$. In the boundary of $\mathcal{P}^*$ an optimal f^* maximizing $l(.)$ exists. Since $f^* = (f_1, \ldots, f_n)$ lies in the convex set $\mathcal{P}^*$ there exist a measure G^* with

$$f_i \;=\; f(x_i|b_i). \tag{6}$$

Since an optimal G^* with not more than n points in the support exists the equations (6) can be solved for G^*. Assuming that θ is fixed means that G^* depends on θ. Therefore the calculation of the posterior distribution needs the application of the above optimization procedure for each θ, $\pi(\theta|D) \propto L(G(\theta)^*)\ \pi(\theta)$. Especially the computation of the posterior mean $\hat{\theta} = \int \theta\pi(\theta|D)\ d\theta$ becomes a troublesome task.

An alternative approach serving as way out is the *marginal method*. We mean that we use a method for calculation that gives an approximate result for the exact optimization procedure described above. The marginal method is much simpler to perform in practice.

The key idea is the interchange of the analyses concerning the parameter of interest θ and the distortion distribution. Here we calculate the marginal distributions first, $f(x_i|b_i)$ is substituted by

$$f(x_i|b_i) \;\simeq\; m(x_i|x_1, \ldots, x_{i-1})b_i \;+\; (1 - b_i) \int g_s(x_i)\ dG(s).$$

In this case the optimization procedure is executed once. Frequently only minor deviations to the exact method occur in practical situations.

The preceding methods remain applicable if the weight α is not observed or missing for observations. In this case we insert α or take the mean of the beliefs, $\bar{b}$ as estimate. Then we end up with

$$f(x_i) \;=\; m(x_i|x_1, \ldots, x_{i-1})\alpha \;+\; (1 - \alpha) \int g_s(x_i)\ dG(s)$$

as model density.

4 Iterative Procedures

The method of convex optimization is accompanied by a useful characterization of the optimum in terms of an equivalence property. Convexity allows the implementation of a directional derivative, here we define

$$\Delta(G_0, G_1) \;:=\; \frac{\partial\, l((1-\epsilon)G_0 + \epsilon G_1)}{\partial\, \epsilon}\Big|_{\epsilon=0}$$

the directional derivative for two measures G_0 and G_1. If the direction G_1 represents a one point measure in s_0, $G_1 = \delta_{s_0}$, then we define

$$\Delta(G_0, s_0) \ := \ \Delta(G_0, G_1).$$

An equivalence theorem (Whittle, 1971, 1973) characterizes the solution of the optimization problem:

The measure G^* maximizes $l(.)$ iff

$$\sup_s \Delta(G^*, s) \ \leq \ \sup_s \Delta(G, s)$$

for all G, or, iff

$$\sup_s \Delta(G^*, s) \ = \ 0.$$

The support of G^* contains only points s_j with $\Delta(G^*, s_j) \ = \ 0$ and consists of not more than n points. This equivalence theorem serves as base of an gradient-based algorithm. Starting at a measure G_0 with k support-points then single points are added successively such that the directional derivative reaches a maximum.

Let G be a measure with a single support point s such that $\Delta(G_0, s)$ is maximized. In the second step we search for a weight ϵ^* with $0 \leq \epsilon^* \leq 1$ such that

$$l((1-\epsilon)G_0 + \epsilon G) \rightarrow \ \max \ .$$

The resulting measure

$$G_1 \ := (1-\epsilon^*)G_0 + \epsilon^* G$$

contains $k+1$ points in the support. For $k \rightarrow \infty$ the sequence of measures G_k converges to the optimal measures G^* in the sense that

$$l(G_k) \ \rightarrow \ l(G^*).$$

It can be shown that convergence holds for any sequence of chosen weights ϵ_k fulfilling the conditions $\epsilon_k \ \rightarrow \ \infty$ and $\sum_k \epsilon_k \ = \ \infty$ (Felsenstein, 1996). The equivalence theorem provides a criterion for stopping the algorithm. In case of optimality the directional derivative should vanish, therefore if

$$\Delta(G_k, s_k^{(i)}) \ \leq \ \delta$$

is fulfilled for a suitable small $\delta > 0$ the iteration might be stopped. Since it suffices to find a measure with at most n points in the support a backwards step after exceeding n points can be added. Therefore points with

$$\Delta(G_k, s_k^{(i)}) \ \geq \ \delta$$

will be eliminated.

The directional derivative for the model with questionable data reads

$$\Delta(G_0, G_1) \ = \ \sum_i \frac{(1-b_i)[\int g_s(x_i) \, dG_1(s) \ - \ \int g_s(x_i) \, dG_0(s)]}{f(x_i|\theta)b_i \ + \ (1-b_i) \int g_s(x_i) \, dG_0(s)}$$

and

$$\Delta(G_0, s_0) \;=\; \sum_i \frac{(1-b_i)[g_{s_0}(x_i) \;-\; \int g_s(x_i)\, dG_0(s)]}{f(x_i|\theta)b_i \;+\; (1-b_i)\int g_s(x_i)\, dG_0(s)}.$$

In the situation where the beliefs are not observable it is possible to apply the same iteration procedure. Then in each step a weight α next to a point s has to be found such that the directional derivative becomes maximum.

If α_0 is the actual weight of G_0 the directional derivative in direction G (with one point s_1 and new weight α) reads

$$\Delta(G_0, s_1, \alpha) \;=\; \sum_i \frac{f(x_i|\theta)[\alpha-\alpha_0] \;+\; (1-\alpha)g_{s_1}(x_i) \;-\; (1-\alpha_0)\int g_s(x_i)\, dG_0(s)}{\alpha_0 f(x_i|\theta) \;+\; (1-\alpha_0)\int g_s(x_i)\, dG_0(s)}.$$

Here the weight α does not vary. The mixture weight is treated as fixed and unknown constant in contrast to a stochastic variable.

5 Example

For purposes of illustration we demonstrate the model with questionable data and its inherent geometry properties in a simple case. Assume the data (if correct) are normally distributed with variance 1 and the prior of the mean is a standard normal distribution. For the distortion distribution we choose a Laplace kernel, therefore

$$g_s(x) \;=\; \frac{1}{2} s\, e^{-|x|s}.$$

Lets compare the marginal method under different assumptions. The meaning and interpretation of the mixture weight is considered here in three ways.

MODEL A: The first model consists of completely specified beliefs for the questionable data. The mixture weight α is a stochastic variable which can be observed. The observations are b_i. This is the model of questionable data we discussed in the previous sections.

MODEL B: Alternatively a fixed but unknown weight α is assumed. Then α is non-stochastic. The optimization procedure involves all possible values of α. The likelihood depends upon the distribution G and the mixture weight α.

MODEL C: A third alternative lies inbetween these two models. Then we assume that the beliefs are provided by the expert but not dedicated to each observation x explicitely. An estimate for α is inserted, for example the mean beliefs $\bar{b}$ serves as estimate of the mixture weight. Then α is non-stochastic, the fixed value is $\bar{b}$.

All these models suit for the concept of convex optimization determining the distortion measure $G(s)$. The assumed prior lead to normal distributions as marginal distributions. Namely the marginal distribution of an single observation is $N(0, 2)$. The covariance of the first and second observation is $cov(x_1, x_2) = 1$. However all conditional marginal distributions are normal in that model and can be calculated without integration due to the conjugate property of the prior.

Assume two observations are taken $x_1 = 1.5$ and $x_2 = -0.4$. The expert trusts in the first observation at 50% and the second 90%. The mean of the beliefs $b_1 = 0.5$, $b_2 = 0.9$ is used as estimate of α in the third model.

The feasible sets are convex and closed subsets of $\mathbb{R}^2$. A point $z = (z_1, z_2) \in \mathbb{R}^2$ belongs to the feasible set if a distortion measure G exists (here that means that a concrete Laplace distribution with parameter s is used) such that

$$z_1 \;=\; f(x_1, b_1)$$

and

$$z_2 \;=\; f(x_2, b_2).$$

The coordinates are the values of the (marginal) densities for the data x_1 and x_2. The feasible set is a curve $(z_1(s), z_2(s))$ depending on the parameter $s \in (0, \infty)$.

The following figure shows the feasible sets for the three models ($F1$ denotes the set for questionable data (Model A), $F2$ denotes the feasible set if the mixture weight is $\bar{b}$ (Model C) and $F3$ denotes the set for Model B.

It includes the optimal points on the boundary of the convex hulls. The optimal points are reached after a couple of steps of the iterative procedure. The determination of the distortion measures simplifies in this situation to the determination of the corresponding s, for the first model that is $s^* = 0.81$. The points $z^* = (z_1(s^*), z_2(s^*)$ maximizing the likelihood are marked in the figure. $\boxtimes$ is the optimal point for model A, $\lhd$ for model B and $\square$ for model C.

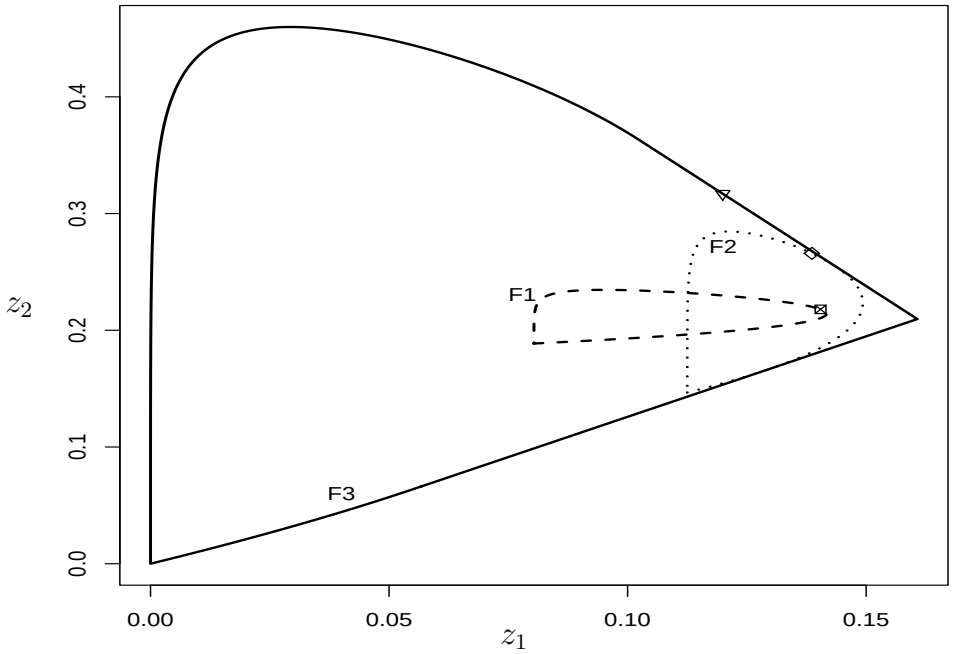

Figure 1: Feasible sets and optimal points

The right vertex in Figure 1 is the point $(m(x_1), m(x_2|x_1))$. This point represents the situation where no doubtful data exist, the belief factor becomes 1. Of course, the

general model ($F3$) includes the case of estimated α ($F2$). The sizes of the sets indicate the flexibility of the model with questionable data.

Under identical circumstances we assume the two observations are $x_1 = 0.35$ with belief $b_1 = 0.72$ and $x_2 = 1.1$ with belief $b_2 = 0.9$. Figure 2 shows the corresponding feasible sets and optimal points. Note that the optimal point of $F2$ belongs to all sets. It may serve as 'compromise'. But for every considered model the point corresponds to its own optimal guess of the distortion measure. In case of questionable data we get

$$f(x) \;=\; m(x|D)b \;+\; (1-b)1.22e^{-|x|/2.44}$$

for these two observations.

The upper vertex indicates the point of complete informative data without questionable data. The optimal point of $F1$ can not be interpreted as mixture within the other models.

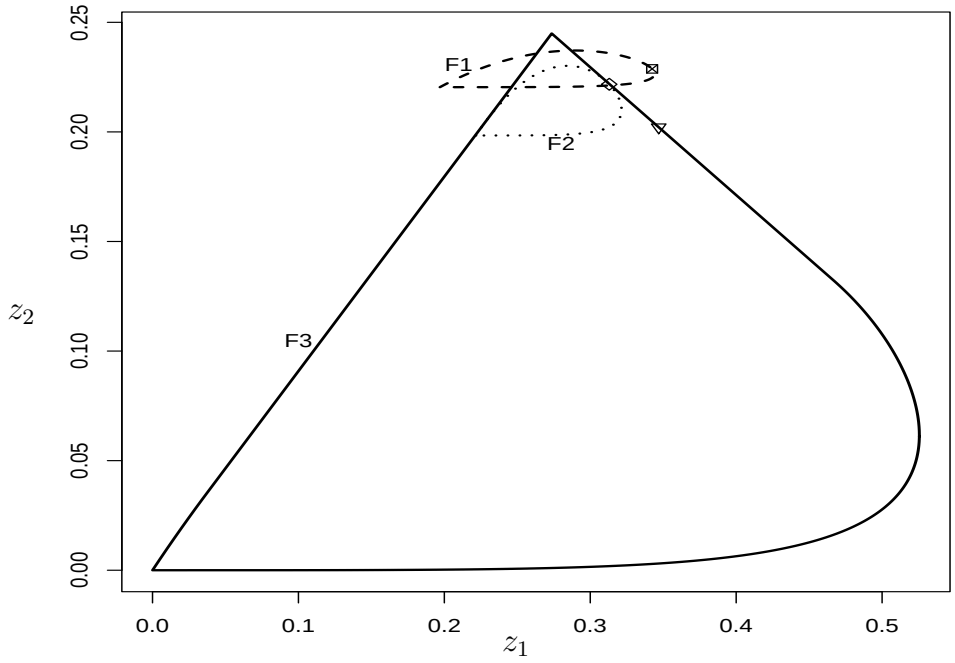

Figure 2: Feasible sets and optimal points

CONCLUDING REMARK: In Molzer et al. (2000) a comprehensive study of road data is presented. A Bayesian regression model is applied to the data. The original regression model turned out to have bad determination. Experts checked the data and completed the data by beliefs. Particularly in that study the additional analysis of beliefs proved useful to improve the quality of the statistical results.

The method can be applied to any data but requires practicable numerical handling. Especially for large datasets the marginal method represents a simplification of calculation. The development of software for questionable data is intended for comfortable usage.

The model is open for several methodical extensions. In the case of prior knowledge about the distribution of distortion it would be sensible to use a nonparametric prior for

the distortion measure. A conjugate Dirichlet process (Antoniak, 1974) e.g. leads to an updating algorithm of the distribution of the distortion measure.

References

C. Antoniak. Mixture of dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, 2:1152–1174, 1974.

M. Bayarri and J. Berger. Robust Bayesian analysis of selection models. *Ann. Statist.*, 25: 645–659, 1998.

M. Bayarri and M. DeGroot. Bayesian analysis of selection models. *The Statistician*, 36: 137–146, 1987.

J. Berger. An overview of Bayesian robustness. *Test*, 3:1–125, 1994.

J. Berger and L. Berliner. Robust Bayes and empirical Bayes analysis with epsilon-contaminated priors. *Ann. Statist.*, 14:461–486, 1986.

J. Berger, B. Betro, E. Moreno, L. Pericchi, F. Ruggeri, G. Salinetti, and L. Wasserman (Eds.). *Bayesian Robustness*, volume 29 of *IMS Lecture Notes*. Springer-Verlag, New York, 1996.

M. DeGroot. A Bayesian view of assessing uncertainty and comparing expert opinion. *J. Statist. Planning and Inf.*, 20:295–306, 1988.

K. Felsenstein. *Bayes'sche Statistik für kontrollierte Experimente*. Vandenhoeck & Ruprecht, Göttingen, 1996.

P. Gustafson and L. Wasserman. Local sensitivity diagnostics for Bayesian inference. *Ann. Statist.*, 23:2153–2167, 1995.

M. Lavine, L. Wasserman, and R. Wolpert. Linearization of Bayesian robustness problems. *J. Statist. Planning and Inf.*, 37:307–316, 1993.

B. Lindsay. *Mixture Models: Theory, Geometry and Applications*. NFS-CBMS Regional Conference Series. IMS, Hayward, 1995.

C. Molzer, K. Felsenstein, R. Viertl, J. Litzka, and A. Vycudil. *Statistische Methoden zur Auswertung von Straßenzustandsdaten*, volume 499 of *Straßenforschung*. BM für Verkehr, Innovation und Technologie, Wien, 2000.

R. Rockafellar. *Convex Analysis*. Princeton University Press, New Jersey, 1972.

F. Ruggeri and L. Wasserman. Infinitesimal sensitivity of posterior distributions. *Canad. J. Statist.*, 21:195–203, 1993.

R. Tibshirani and L. Wasserman. Some aspects of the reparameterization of statistical models. *Canad. J. Statist.*, 22:163–173, 1994.

P. Whittle. *Optimization under Constraints*. Wiley, London, 1971.

P. Whittle. Some general points in the theory of optimal experimental design. *J. Royal Statist. Soc. B*, 35:123–130, 1973.

Author's address:

Klaus Felsenstein
Department of Statistics and Probability Theory
Vienna University of Technology
Wiedner Hauptstr. 8-10
A-1040 Vienna

Tel. +43 1 58801 / 10722
E-mail: Klaus.Felsenstein@tuwien.ac.at