Profile Methods

Christian Ritter[†] and Douglas M. Bates[‡]
[†]CORE, Université Catholique de Louvain
34, voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium
[‡]Department of Statistics, University of Wisconsin
1210 W Dayton Street, Madison, WI 53706, USA.

February 23, 1994

Abstract

We begin by describing profile methods, such as profile values, profile traces, profile transformations, and profile diagnostic plots, in a general setting and point out an interesting feature of profile transforms, the boxing property. We then discuss profile methods in the specific contexts of likelihood and Bayesian inference and make links with approximate Bayesian marginalization, with Jeffrey's prior, and with importance sampling. We conclude by investigating profile transforms for posterior distributions and point out some peculiarities.

Keywords: Bayesian inference, boxing property, likelihood, nonlinear regression.

1 Introduction

In this article we introduce profile methods and discuss some of their uses. We begin with a general mathematical setting and present several examples. Later, we describe how profile techniques arise in statistical inference, where we distinguish between likelihood and Bayesian treatments.

In statistics, the term *profile* first appeared in the literature on likelihood inference in the presence of nuisance parameters (Barndorff-Nielsen, 1983; Barndorff-Nielsen, 1986; Cox and Reid, 1987; McCullagh and Tibshirani, 1990). Since then, it has been used in the contexts of nonlinear regression and Bayesian statistics. Bates and Watts (1988) used profiles of the sum of squares function for likelihood inference for nonlinear models; Hills and Smith (1993) used profiles of posterior distributions to assess the effectiveness of parameter transformations aimed at increasing *a posteriori* "normality."

2 General Profile Methods

Profiling methods are very general tools for summarizing a function F of several variables $\theta = (\theta_1, \ldots, \theta_p)'$ in a neighborhood of an extremum $\hat{\theta}$. Each parameter component θ_i is varied in turn, and for each value $\theta_i = t$ the function F is re-optimized with respect to the other components. The location of the conditional optimum

$$\tilde{\theta}^{(i)}(t) := \arg\text{opt}\,\{F(\theta)|\theta_i = t\}$$

taken as a function of t, is called the ith profile trace of F, and $\tilde{F}_i(t) = F(\tilde{\theta}^{(i)}(t))$ is called the ith profile value. Accordingly, we call the structure $\left\{\tilde{\theta}^{(i)}(\cdot), \tilde{F}_i(\cdot)\right\}$ the ith profile of F. If one pictures F as a mountain over the parameter domain, then the ith profile value is the profile of the mountain when seen from the θ_i axis. If the Hessian of F is non-singular in $\hat{\theta}$, the existence and uniqueness of all profiles in a neighborhood of $\hat{\theta}$ is guaranteed by the implicit function theorem. Further away, the conditional optimum may no longer exist or may not be unique.

In practice, profiles are computed at discrete points, beginning from the center $\hat{\theta}$ and going outward in each parameter component until either a desired range has been covered, or the optimization routine fails to converge. Intermediate values of the profiles can be obtained by spline interpolation.

Example 1: Suppose $F(\theta) = \theta' A \theta$ with

$$A = \begin{pmatrix} \frac{4}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{4}{3} \end{pmatrix}.$$

Suppose also that $\eta_1 = \exp(\theta_1)$ and $\eta_2 = \exp(\theta_2)$. Figure 1 shows the resulting level contours of F for the levels 0.25, 1.0, and 2.25 and the profile traces for F both in the θ and the η coordinates. Since the profile traces are the conditional optima (here, minima), contours of F must intersect the first profile traces vertically and the second profile traces horizontally.

Figure 2 displays the corresponding profile values for θ_1 and η_1. Since F is symmetric in both the θ and the η coordinates, the profile value function for θ_2

looks the same as the one for θ_1 and the profile value function for η_2 looks the same as the one for η_1.

Another, often more effective, way of displaying the profile values is by plotting the functions

$$\zeta_i(t) = \operatorname{sgn}(t - \hat{\theta}_i)\left[\tilde{F}_i(t) - \hat{F}\right]^{\frac{1}{2}} .$$

where $\hat{F}$ is the global minimum. If F is exactly quadratic, then the functions $\zeta_i(t)$ are linear. The resulting plots for θ_1 and η_1 are shown in Figure 3. The ζ_1 value at $t = \hat{\theta}_1$ is zero, and the slope at $\hat{\theta}_1$ is equal to 1, which is the square-root of the first diagonal element of A^{-1}.

In general, we define the profile-ζ functions by

$$\zeta_i(t) = \begin{cases} \operatorname{sgn}(t - \hat{\theta}_i)\left[\tilde{F}_i(t) - \hat{F}\right]^{\frac{1}{2}}, & \hat{\theta} \text{ minimum of } F \\ \operatorname{sgn}(t - \hat{\theta}_i)\left[\hat{F} - \tilde{F}_i(t)\right]^{\frac{1}{2}}, & \hat{\theta} \text{ maximum of } F \end{cases}$$

When the profile-ζ functions are monotone, they define a set of one-to-one mappings $\theta_i \mapsto \zeta_i$, which we call the profile-ζ transformations. In Example 1, the profile transformations become $\zeta_i(\theta_i) = \log(\theta_i)$, $i = 1, 2$. For this example the profile transformations define a system of coordinates in which F is quadratic.

This leads to an important diagnostic plot in which the profile-ζ functions are displayed together with the pairwise projections of the profile traces in the original and in the ζ coordinates. Figure 4 shows this plot for Example 2.1 in the η coordinates; it has been enhanced by adding selected contours of F.

Suppose F is already quadratic in the θ coordinates. Then F is again quadratic in the ζ coordinates and it has the standardized form $F(\zeta) = \hat{F} + \zeta' C^{-1} \zeta$, with the matrix

$$C = D^{-1} S D^{-1}$$

where

$$D = diag\left(|S_{11}|^{\frac{1}{2}}, \cdots, |S_{pp}|^{\frac{1}{2}}\right)$$

and $S = \{S_{ij}\}$ is the inverse of the Hessian of F evaluated at $\hat{\eta}$. This implies that F is a standard parabola $\hat{F} \pm \zeta_i^2$ along the ζ_i axes and that the profile traces in the ζ coordinates are straight.

This remains true in many cases where F is given in a system of coordinates which is the result of component-wise transformations of another system of coordinates in which F is quadratic, such as in Example 1.

If the coordinates F is given in are not the result of such transformations, F is still of the form of a standard parabola along each ζ_i axis, but the profile traces in the ζ coordinates are curved. This limited regularization is shown in Example 2.

Example 2: Suppose that F is defined as in Example 1 and that it is given to us in the coordinates ϕ_1 and ϕ_2 defined by $\theta_1 = \sinh(0.5(\phi_1 + \phi_2))$ and $\theta_2 = \exp(\phi_1) - \exp(\phi_2)$. Figure 5 shows the resulting profile diagnostic plot. Clearly, the profile-ζ transforms were not successful at making F quadratic. Nevertheless, level contours of F in the ζ coordinates are bounded by concentric squares, which is not true in the original ϕ coordinates. This property of the profile transformations remains valid in the general case; contours of level $\hat{F} + r$ (if $\hat{F}$ is the minimum) or $\hat{F} - r$ (if $\hat{F}$ is the maximum) are bounded by concentric hypercubes (boxes) $[-r^{\frac{1}{2}}, r^{\frac{1}{2}}]^p$. We call this the boxing property.

3 Profile Methods in Likelihood Inference

3.1 General

Likelihood inference about the parameter vector θ is based on the likelihood function $L(\theta|x)$. Joint likelihood regions are bounded by level contours of L, and their frequency coverage can be calculated using the χ^2 approximation to the log-likelihood ratio statistic $2\left[\ln L(\hat{\theta}|x) - \ln L(\theta|x)\right]$.

For linear models with normal errors, likelihood regions are concentric ellipsoids and their calculated coverage levels are exact. For nonlinear models, likelihood regions are usually neither concentric nor elliptical. The simplest approximate likelihood inference for nonlinear models is based on a normal distribution with mean vector equal to the maximum likelihood estimates and covariance matrix equal to the inverse of the observed information matrix. This is acceptable as long as the level contours of the likelihood can be approximated reasonably well by concentric ellipses. Especially in small sample applications this approximation may however not be adequate.

There are two important uses of profiles in this context. First, the profile traces and the profile-ζ functions can be plotted; that is, the profile diagnostic plot defined in the previous section can be constructed. If the profile traces and the profile-ζ functions are all straight, the simple normal approximation is likely good. Second, one can use the profile-ζ transforms defined in the previous section to make the log-likelihood appear more quadratic. Then one can conduct simple approximate likelihood inference in the transformed coordinates. The profile transformations depend on the data, but this dependence is usually mild.

3.2 Nonlinear Regression

The use of profile transforms in nonlinear regression inference has been developed by Bates and Watts (1988). To avoid confusion with the general profile-ζ transforms, we call them the profile-t transforms. In order to normalize the sum of squares $S(\beta)$ and to make it appear more quadratic, Bates and Watts use

$$\tau_i(t) = \operatorname{sgn}(t - \hat{\theta}_i) \left\{ [\tilde{S}_i(t) - S(\hat{\theta})]/s^2 \right\}^{\frac{1}{2}}$$

where s^2 is an estimate of the variance σ^2 based on q degrees of freedom. They then employ the approximate F statistic $[S(\tau) - \hat{S}]/(p \cdot s^2)$ to calculate coverage levels for likelihood regions in the new coordinates.

A simple technique for deriving such approximate inference regions is to construct ellipsoidal approximations to the level contours of the sum of squares function in the τ coordinates. The pairwise projections of these ellipsoids can then be displayed directly in the τ coordinates or they can be back-transformed and displayed in the θ coordinates. Bates and Watts (1988), Appendix A6, go one step further and incorporate the information from the profile traces in an interpolation algorithm for approximating the pairwise projections of the level contours of $S(\tau)$. In many practical situations, their method produces very good approximations to the true pairwise projections of the contours of S in the τ as well as in the θ coordinates. These approximate contours can be incorporated into the pairwise profile diagnostic plot.

3.3 General Likelihoods

The same technique, now with profile transforms based on the log-likelihood and on the χ^2 approximation to the log-likelihood ratio, can be used to generate approximate pairwise projections of general joint likelihood regions. If, in Example 2, we treat $-F(\psi_1, \psi_2)$ as a log-likelihood, we can compute its exact and approximate level contours both in the ψ and in the ζ coordinates. Figure 6 shows the resulting pairwise plot augmented by the approximate contours (dotted) and the exact ones (solid). The exact contours have been computed by evaluating L over a fine grid. Grid-based contouring is still feasible for two dimensions, but the required computing effort grows exponentially with the number of parameters. On the other hand, since the computing effort for obtaining the profile-based approximate contours is roughly of polynomial order in the number of parameters p, such contours can require substantially less computing even for values of p as small as 3 or 4.

3.4 Marginal Inference

Often, some parameters are of more interest than others. In this case, the likelihood has to be "marginalized." There is no consensus on how this can best be done. Common strategies are: simple profile likelihoods, modified profile likelihoods (Barndorff-Nielsen, 1983; Barndorff-Nielsen, 1986; McCullagh and Tibshirani, 1990), conditional orthogonalized likelihoods (Cox and Reid, 1987), and integrated likelihoods (Kalbfleisch and Sprott, 1970). The latter method is close to Bayesian inference, since before integrating out the nuisance parameters, vague priors are defined.

The simplest form of the marginal of the ith parameter is the direct profile likelihood $\tilde{L}_i(t)$. It supresses the fact that the conditional maximum $\tilde{\theta}^{(i)}(t)$ is also subject to variation; inference based on it tends to underestimate the variance of θ_i. In extreme cases, such as the many means problem described in Cox and Reid (1987), this produces inconsistent estimators. If $\tilde{\theta}^{(-i)}(t)$ denotes $\tilde{\theta}^{(i)}(t)$ without its ith component, the quantity

$$\tilde{\sigma}_{-i}(t) = |\tilde{R}_{-i}(t)|^{-\frac{1}{2}}$$

where

$$\tilde{R}_{-i}(t) = \frac{\partial^2}{\partial\theta_{-i}\partial\theta'_{-i}} \log L(\theta)\bigg|_{\theta = \tilde{\theta}^{(i)}(t)}.$$

is an approximate measure of variability of $\tilde{\theta}^{(-i)}(t)$. Values of θ_i are more "likely" if the variability $\tilde{\sigma}_{-i}(\theta_i)$ is large, since then many realizations of $\theta^{(-i)}$ are compatible with θ_i. The adjusted marginal profile likelihood $\tilde{L}_i(t) \cdot \tilde{\sigma}_{-i}(t)$ makes the appropriate correction. It is similar to the modified profile likelihood proposed by Barndorff-Nielsen (1983).

3.5 Marginal Likelihoods in Nonlinear Regression

In the nonlinear regression model $y_i \sim \mathcal{N}(f(\theta, x_i), \sigma^2),\ i = 1, \ldots, n$, the full likelihood is proportional to $(\sigma^2)^{-\frac{1}{2}n} \cdot \exp\left(-\frac{1}{\sigma^2}S(\theta)\right)$. If σ^2 is eliminated by simple profile likelihood, that is, by substituting $\hat{\sigma}^2 = \frac{1}{n}S(\theta)$, the resulting reduced likelihood is proportional to $S(\theta)^{-\frac{1}{2}n}$; whereas, if the modified profile likelihood is used instead, it is $S(\theta)^{-(\frac{1}{2}n-1)}$.

Both expressions also correspond to integrated likelihoods; in the first case, a flat prior is put on $\log \sigma^2$; in the second, on σ^2. Since a log-constant prior for the noise level σ^2 is usually considered a better choice we limit our attention to $L(\theta) \propto S(\theta)^{-\frac{1}{2}n}$. The reduced likelihood $L(\theta)$ has the form of a generalized multivariate-t distribution. In linear regression with design matrix X, it reduces to an exact multivariate-t distribution with location vector $\hat{\theta}$ and scale matrix $\hat{\sigma}^2(X'X)^{-1}$ on $n-p$ degrees of freedom (Box and Tiao, 1973, Section 2.7.2).

For likelihood inference about a component of θ, we can choose among several approaches. The simple profile likelihood is

$$\tilde{L}_i(\theta_i) = \tilde{S}_i(\theta_i)^{-\frac{1}{2}n},$$

while the modified profile likelihood is

$$L(\theta_i) = \tilde{S}_i(\theta_i)^{-\frac{1}{2}n} |\tilde{R}_{-i}(\theta_i)|^{-\frac{1}{2}}.$$

The Hessian $\tilde{R}_{-i}(\theta)$ can also be written as $\tilde{S}_i(\theta_i)\tilde{U}_{-i}(\theta_i)$ with

$$\tilde{U}_{-i}(t) = \left. \frac{\partial^2}{\partial\theta_{-i}\partial\theta'_{-i}} S(\theta) \right|_{\theta=\tilde{\theta}^{(i)}(t)}.$$

Thus, the marginal likelihood becomes

$$L(\theta_i) = \tilde{S}_i(\theta_i)^{-\frac{n-p+1}{2}} \cdot |\tilde{U}_{-i}(\theta_i)|^{-\frac{1}{2}}.$$

This form of the marginal likelihood displays a width adjustment and a correction of the degrees of freedom. In linear regression, this degree-of-freedom adjustment reduces the likelihood exactly to a t_{n-p} distribution, which is equal to the likelihood of the corresponding frequency statistic.

The determinant of the Hessian $\tilde{U}_{-i}$ can be obtained as a by-product of whichever algorithm generates the profiles of S, such as the common Gauss-Newton scheme.

3.6 Confidence Statements About Functionals of the Parameters

Another common task is making confidence statements about functions $\rho = g(\theta)$ of the original parameters. It is not obvious how to do this in the likelihood framework. One strategy is to reparameterize the problem such that

$(\rho, \gamma)' = (g(\theta), H(\theta))$ is a regular one-to-one transformation, and then use the techniques we described before to find the marginal likelihood for ρ. Another strategy is described in Clarke (1987) for nonlinear least squares, where a reparameterization $\theta \mapsto \phi$ is sought, in which the sum of squares is approximately quadratic. In the transformed coordinates, confidence regions based on the log likelihood-ratio are well-approximated by ellipsoids. Under some regularity conditions on the function g, approximate confidence intervals of level $1-\alpha$ can then be bounded by the minimum and the maximum of $g(\phi)$ on the ellipse defined by $2 \cdot \left[\log L(\hat{\phi}) - \log L(\phi)\right] = \chi^2_1(1-\alpha)$. Clarke recommends constructing transformations based on minimum bias, but the one-to-one profile transforms can also make the sum of squares function or the log-likelihood more quadratic.

4 Profile Transformations for Bayesian Inference Under Flat Priors

Let us assume that we are either working with a general likelihood or with a nonlinear regression likelihood proportional to $S(\theta)^{-n/2}$ (that is, after eliminating σ^2). Under favorable conditions, the integral of the likelihood $L(\theta)$ is finite, and therefore, $L(\theta)$ can be normalized and treated as a posterior. If the integral is infinite, either the parameter space can be truncated, or a regularizing but sufficiently vague prior can be sought.

In nonlinear regression, treating $S(\theta)^{-\frac{1}{2}n}$ as a probability density is an intuitive way of measuring uncertainty about the parameters. Parameter vectors with small sums of squares are more likely than competitors with larger sums of squares, and it seems reasonable to choose their probabilities proportional to the likelihood.

Bayesian inference from likelihoods under flat priors includes posterior modes, i.e., maximum likelihood estimates, posterior means, i.e. , integrals of the likelihood, and HPD regions, i.e. regions bounded by level contours of the likelihood. The conclusions drawn from such inferences are similar or even identical to those drawn using the frequentistic techniques described earlier. Nevertheless, there are some important distinctions.

In Bayesian inference, the choice of the prior distribution matters. Treating the likelihood in the θ coordinates as a posterior is not the same as doing so in the transformed ζ coordinates. In the first case, a flat prior is placed on θ; in the second case, on ζ, the set of transformed parameters. Going from the θ to the ζ

coordinates, or vice versa, requires the Jacobian of the parameter transformation.

A flat prior on θ means that without knowledge of the data, all parameter vectors θ are equally possible. This is not always realistic. Jeffrey's (Box and Tiao, 1973, Section 1.3.5) idea of using the square root of the determinant of the local information matrix as a prior may be a cleaner, transformation invariant, technique for converting likelihoods into posteriors; yet, using such a prior requires more computing effort. If the log-posterior is quadratic or nearly quadratic, Jeffrey's prior is flat or approximately flat. Thus, if the profile transforms fulfill their purpose of making the log-likelihood look quadratic, using a flat prior in the transformed coordinates is roughly equivalent to using Jeffrey's prior in the original coordinates. In addition, using a flat prior in the transformed coordinates is a reasonable way of expressing prior uncertainty (Box and Tiao, 1973, Section 1.3.5).

Furthermore, the likelihood in the transformed coordinates is often integrable even if the likelihood in the original coordinates is not. For nonlinear regression, the likelihood $S(\tau)^{-\frac{1}{2}n}$ in the transformed coodinates is integrable if the profiles exist and are unique over the entire parameter space $\Theta = R^p$, and if each of the profile-t transforms is $1-1$ from R to R. In this case, the transformed parameter space is again R^p, and the boxing property holds globally. The latter assures that $S(\tau) \geq S(0) + ||\tau||^2/p$, and therefore,

$$[S(\tau)]^{-\frac{1}{2}n} \leq \left[S(0) + \frac{||\tau||^2}{p}\right]^{-\frac{1}{2}n}.$$

The right side of the equation is proportional to a multivariate-t density with $n-p$ degrees of freedom and scale matrix $p \cdot s^2 \cdot I_p$, where $s^2 = S(0)/(n-p)$, and I_p is the p-dimensional identity matrix. This assures, for example, that numerical integration by importance sampling based on a circular multivariate-t distribution with at most $n-p$ degrees of freedom will be safe from problems in the tails.

5 Profiling and Approximate Marginals

Laplacian approximations (De Bruijn, 1961) of Bayesian marginals of functionals $\eta(\theta)$ are typically of the form

$$\pi^*_{x;\eta}(t) \propto \tilde{\pi}_{(x;\eta)}(t) \cdot s_\eta(t, \pi),$$

where $\tilde{\pi}_{(x;\eta)}(t) = \max\{\pi_x(\theta)|\eta(\theta) = t\}$ and $s_\eta(t,\pi)$ is an adjustment factor depending on t through $\tilde{\theta}^{(\eta)}(t)$, which is a conditional maximizer of $\pi_x(\theta)$ subject to $\eta(\theta) = t$. Key references are Leonard (1982), Tierney and Kadane (1986), Tierney, Kass and Kadane (1989), and Leonard, Hsu and Tsui (1989).

In the special case where $\eta(\theta) = \theta_i$ the adjustment factor is simply $|\tilde{R}_{-i}(\eta)|^{-\frac{1}{2}}$, the $(p-1)\times(p-1)$ Hessian matrix of $-\log\pi(\theta)$ with respect to θ_{-i} evaluated at $\tilde{\theta}^{(i)}(t)$. If the i-th profile $\left\{\tilde{\theta}^{(i)}(\cdot); \tilde{\pi}_{(x;i)}(\cdot)\right\}$ exists and is unique for all values of θ_i, we only need to compute $|R_{-i}(\cdot)|^{\frac{1}{2}}$ along the profile $\tilde{\theta}^{(i)}(\cdot)$, to permit evaluation of $\pi^*_{x;i}(\theta_i)$. In some situations, this is straightforward.

If $\pi_x(\theta)$ is a least squares posterior, i.e., $\pi_x(\theta) \propto S(\theta)^{-\frac{1}{2}\nu}$, where $S(\theta)$ is a sum of squares, then the conditional maximizations can be accomplished by second order algorithms, such as the Gauss-Newton method, which evaluate $\tilde{R}_{-i}(\theta_i)$ as a by-product of computing $\tilde{\theta}^{(i)}(\theta_i)$. Thus, for those posterior distributions, the profiles can routinely be augmented to $\left\{\tilde{\theta}^{(i)}(\cdot); \tilde{\pi}_{(x;i)}(\cdot); |\tilde{R}_{-i}(\cdot)|^{\frac{1}{2}}\right\}$, making the approximate marginals of the components easily accessible.

If we are not working with a least squares posterior we may still be able to obtain the profile $\left\{\tilde{\theta}^{(i)}(\cdot); \tilde{\pi}_{(x;i)}(\cdot)\right\}$. However, these conditional optimizations are commonly carried out using quasi-Newton algorithms which, in general, do not produce good estimates of the Hessian matrix at the optimum.

If $\pi_x(\theta)$ is of a sufficiently simple form, it may be possible to compute $R_{-i}(\cdot)$ analytically; this has become more feasible due to recent advances in mathematical software. Otherwise, one can compute numerical Hessians along the profile traces. This can be done by evaluating $\pi_x(\theta)$ over Koshal designs (Koshal, 1933) centered at $\tilde{\theta}^{(i)}(\theta_i)$. The computing effort for this technique is on the order of $\frac{1}{2}p^2 + p$ evaluations of $\pi_x(\theta)$, where p is the dimension of θ. It is practical for problems of small or moderate dimension p. It may be possible to develop more efficient generic methods if one exploits the fact that only $|\tilde{R}_{-i}|$ but not $\tilde{R}_{-i}$ itself is needed.

6 Profile Transformations For Posteriors

Suppose that the p-dimensional vector θ with domain Θ has likelihood $L(\theta|x)$ and prior density $\pi(\theta)$; the logarithm $h(\theta)$ of the posterior $\pi_x(\theta) \propto L(\theta|x)\pi(\theta)$ exists on Θ and has a maximum θ^* as a solution to the equations $\nabla h(\theta) = 0$; and the Hessian R^* of $-h(\cdot)$ evaluated at $\theta^*$is nonsingular. Following the arguments of Section 2, the (posterior) profiles $\left\{\tilde{\pi}_{(x;i)}(\cdot), \tilde{h}_i(\cdot)\right\}$ exist and are unique in a neighborhood of θ^*.

Now suppose that, *a posteriori*, θ has an exact multivariate normal distribution with mean vector μ and covariance matrix Σ. Then the profile traces $\tilde{\theta}^{(i)}(t)$ are linear, and the profile-ζ functions

$$\zeta_i(t) = \operatorname{sgn}(t - \theta_i^*) \left\{ -2 \left[\tilde{h}_i(t) - h(\theta^*) \right] \right\}^{\frac{1}{2}}$$

are straight lines in t. Moreover, the profile traces $\tilde{\zeta}^{(i)}(\cdot)$ in the transformed coordinates will be equal to the transformed traces $T[\tilde{\theta}^{(i)}(\cdot)]$, and the profile values $g_{(i)}(s) = g(\tilde{\zeta}^{(i)}(s))$, with g as the log posterior in the ζ coordinates, will be equal to $h(T^{-1}(\tilde{\zeta}^{(i)}(s)))$ up to an additive constant. The posterior in the transformed parameters is a multivariate normal distribution with mean vector zero and covariance matrix C, where C is the correlation matrix corresponding to Σ.

If θ does not have a multivariate normal distribution, the functions $\zeta_i(\theta_i)$ will no longer be linear, and therefore, the Jacobian of T has to be taken into account. Let us denote the posterior in the transformed coordinates by $q(\zeta)$. The traces $\tilde{\zeta}^{(i)}(\cdot)$ of $q(\zeta)$ are not equal to the transformed original traces, since they are now based on $\pi_x[T^{-1}(\zeta)]/|\dot{T}[T^{-1}(\zeta)]|$ and not on $\pi_x[T^{-1}(\zeta)]$. Moreover, along the traces, the profile values depend on the Jacobian via $\tilde{g}_i(s) = h\{T^{-1}[\tilde{\zeta}^{(i)}(s)]\} - \ln\{|\dot{T}[T^{-1}(\tilde{\zeta}^{(i)}(s))]|\}$ and therefore, in general, are not quadratic in s. However, the difference will be small unless the transformation T is strongly distorting.

Hills and Smith (1993) suggest using profile diagnostic plots of the posterior to assess "normality". If this assessment is negative, they suggest applying a set of transformations of the parameters, to recompute the profiles, and to re-examine the diagnostic plots. The shape of the profile-ζ values can help in choosing such transformations and Hills and Smith recommend using a lookup table of profile-ζ values and appropriate transformation to facilitate this task.

The top plot in Figure 7 shows the transformed profile traces $T[\tilde{\theta}^{(i)}(\cdot)]$ and the profile traces $\tilde{\zeta}^{(i)}(\cdot)$ in the transformed coordinates for Example 2, while the bottom plot shows the contours of the transformed posterior $p[T(\zeta)]$ and the posterior $q(\zeta)$ in the transformed coordinates. Figure 8 shows the profile values in the ζ coordinates, with and without consideration of the Jacobian.

If one is willing to accept the transformed profile traces as approximations to the profile traces in the transformed coordinates, one can attempt to improve the degree to which the log-posterior in the new coordinates is quadratic by the following procedure.

1. Compute the profiles $\left\{\tilde{h}_i(\cdot), \tilde{\theta}^{(i)}(\cdot)\right\}$ and the simple transformation $T_0 \colon \theta \mapsto \zeta$

by

$$\tilde{\zeta}_i(t) = \mathrm{sgn}(t - \theta_i^*)\left\{-2[\tilde{h}_i(t) - h^*]\right\}^{\frac{1}{2}}$$

2. For $k = 0, 1, \ldots$ iterate:

   - define $\breve{h}_i(\cdot)$ as $\breve{h}_i(t) = \tilde{h}_i(t) - \ln|\dot{T}_k[\tilde{\theta}^{(i)}(t)]|$
   - compute $\breve{\theta}_i = \arg\min[-\breve{h}_i(t)]$
   - define $T_{k+1}\colon \theta \mapsto \zeta$ by

$$\tilde{\zeta}_i(t) = \mathrm{sgn}(t - \breve{\theta}_i^*)\left\{-2[\breve{h}_i(t) - \breve{h}_i(\breve{\theta}_i)]\right\}^{\frac{1}{2}}$$

One iteration of the above algorithm may be sufficient. Note that the algorithm may not always be feasible, since $-\breve{h}_i(\cdot)$ need not have a unique minimum, or any minimum at all.

For our example, Figure 9 shows the profile values in the ζ coordinates, with and without consideration of the Jacobian after one iteration of the adjustment algorithm.

Further improvement can be achieved by reevaluating the posterior profile traces in the new parameters. In many cases, the transformed profile traces will be close to the profile traces in the new parameters, and therefore, the transformed traces will provide good starting estimates for the conditional optimization procedures.

Instead of the profile transformations based on the posterior, one can try to use the ones based on the likelihood. If the likelihood dominates the prior distribution, the profile transformations derived from the likelihood and from the posterior are very similar. For many practical situations, like nonlinear regression, profile transforms are easier to compute from the likelihood than from the posterior. Again, one can adjust the transformation by the iterative procedure outlined in a previous paragraph.

Genuine posterior profiles are the only option in situations where profiles of the likelihood cannot be computed because there is no unique maximum likelihood estimates, but where the prior distribution is strong enough to enforce a unique mode of the posterior distribution. The ESCA problem, discussed in detail in Ritter (1993), is an example of such a situation.

7 Further Remarks on Profiling Methods

Profiles, whose computation is generic and can be automated, are natural ingredients in a wide range of inference tasks. Until a short time ago, the computing effort for profiling had been prohibitive for routine analyses, but recent advances in computing power and improved algorithms have made profiling feasible for many difficult problems of small and moderate dimensions. The user has to be able to formulate the model and to interpret the results and diagnostics. The latter requires some training and experience but is not model specific.

In practice, one needs to distinguish between profiling methods for nonlinear regression and profiling methods for general likelihoods and posteriors. For nonlinear regression, the optimizations are much easier than for general objective functions, and the usual algorithms automatically provide interesting additional information like conditional Hessians. As computing power continues to grow and algorithms for symbolic differentiation become more widely available and user-friendly, there will be fewer obstacles to the routine use of profile methods for more general likelihoods and posteriors. First generic implementations of profile methods can be found in the S language (Chambers and Hastie, 1992) in form of the functions profile.nls, for nonlinear regression, and profile.ms, for general likelihoods; user implementations are straightforward in other statistical languages which contain nonlinear least squares and general minimization modules and which have sufficiently flexible command languages.

8 Acknowledgements

This work was partly funded by the National Science Foundation (USA) and by the Human Capital Program of the European Community.

References

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator, *Biometrika* **70**: 343–365.

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

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*, John Wiley, New York.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference In Statistical Analysis*, Addison Wesley Publishing, Reading, MA.

Chambers, J. M. and Hastie, T. J. (eds) (1992). *Statistical Models in S*, Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, California.

Clarke, G. P. Y. (1987). Approximate confidence limits for a parameter function in nonlinear regression, *Journal of the American Statistical Association* **82**: 221–230.

Cox, D. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference, *Journal of the Royal Statistical Society Series B* **49**: 1–18.

De Bruijn, N. G. (1961). *Asymptotic Methods in Analysis*, North-Holland, Amsterdam.

Hills, S. E. and Smith, A. F. M. (1993). Diagnostic plots for improved parametrization in Bayesian inference, *Biometrika* **80**: 61–74.

Kalbfleisch, J. and Sprott, D. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion), *J. R. Statist. Soc. B* **32**: 175–208.

Koshal, R. (1933). Application of the method of maximum likelihood to the improvement of curves fitted by the method of moments, *Journal of the Royal Statistical Society Series A* **96**: 303–313.

Leonard, T. (1982). Comments on “A simple predictive density function”, *Journal of the American Statistical Association* **77**: 657–658.

Leonard, T., Hsu, J. S. J. and Tsui, K. W. (1989). Bayesian marginal inference, *Journal of the American Statistical Association* **84**: 1051–1058.

McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods, *Journal of the Royal Statistical Society Series B* **52(2)**: 325–344.

Ritter, C. (1993). Statistical analysis of ESCA spectra, *The Statistician*. submitted.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and densities, *Journal of the American Statistical Association* **81**: 82–86.

Tierney, L., Kass, R. and Kadane, J. B. (1989). Approximate marginal densities of nonlinear functions, *Biometrika* **76**: 425–433.

List of Figures

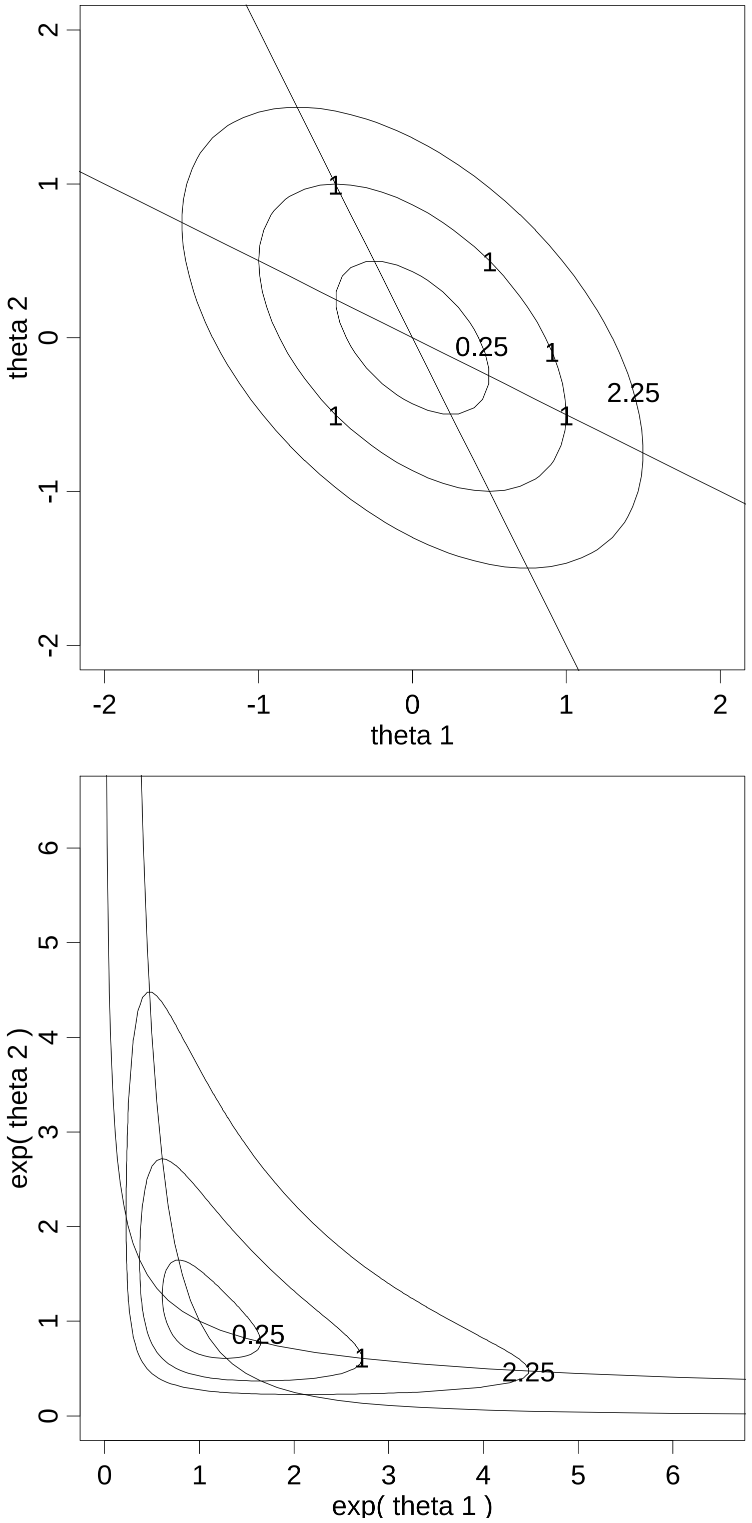


Figure 1: Contours and profile traces of F in θ and η coordinates

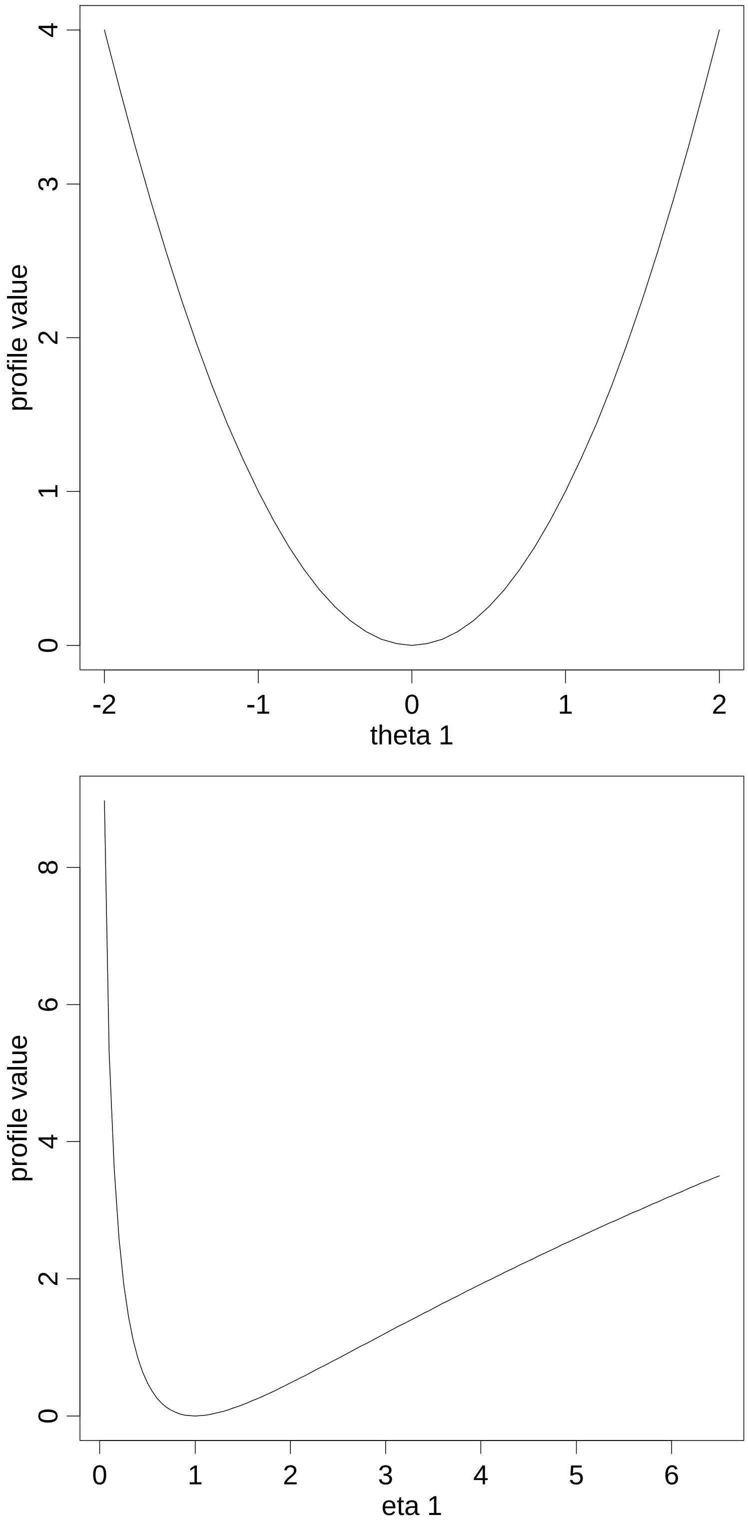


Figure 2: Profile value function of F with respect to θ_1 and η_1

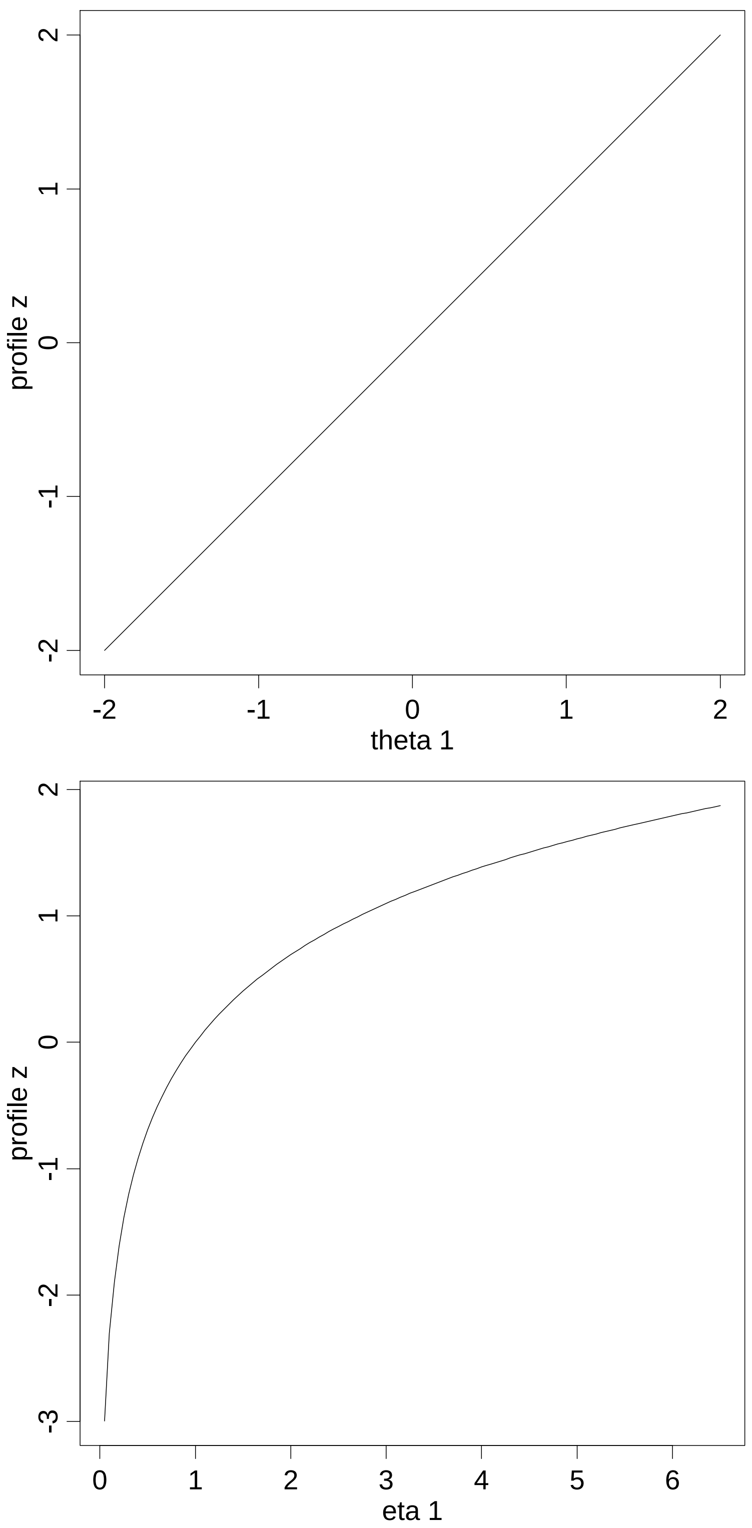


Figure 3: Profile-ζ plot in θ and η coordinates

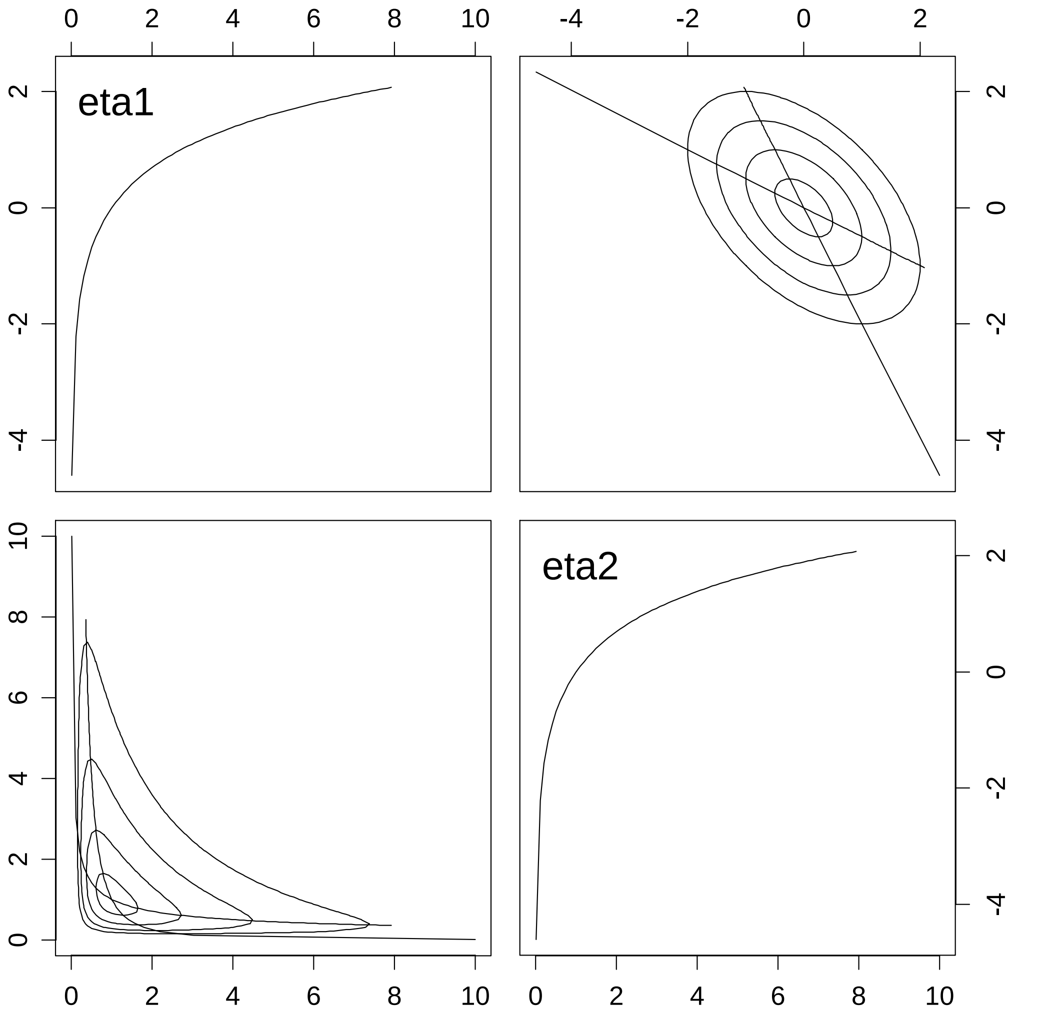


Figure 4: Profile pairs diagnostic plot for Example 1. Upper left: ζ_1 versus η_1; upper right: ζ_1 versus ζ_2; lower left η_2 versus η_1; lower right: ζ_2 versus η_2.

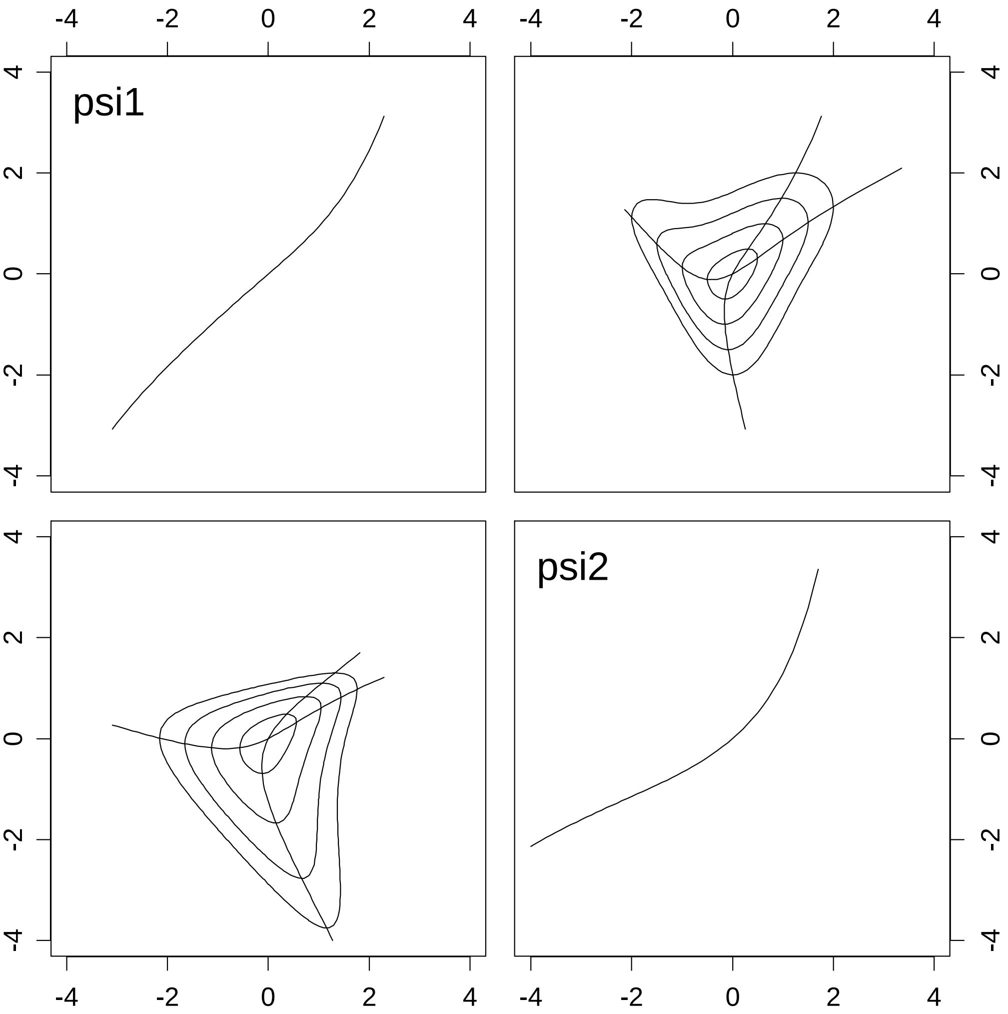


Figure 5: Profile pairs diagnostic plot for Example 2. Upper left: ζ_1 versus ψ_1; upper right: ζ_1 versus ζ_2; lower left ψ_2 versus ψ_1; lower right: ζ_2 versus ψ_2.

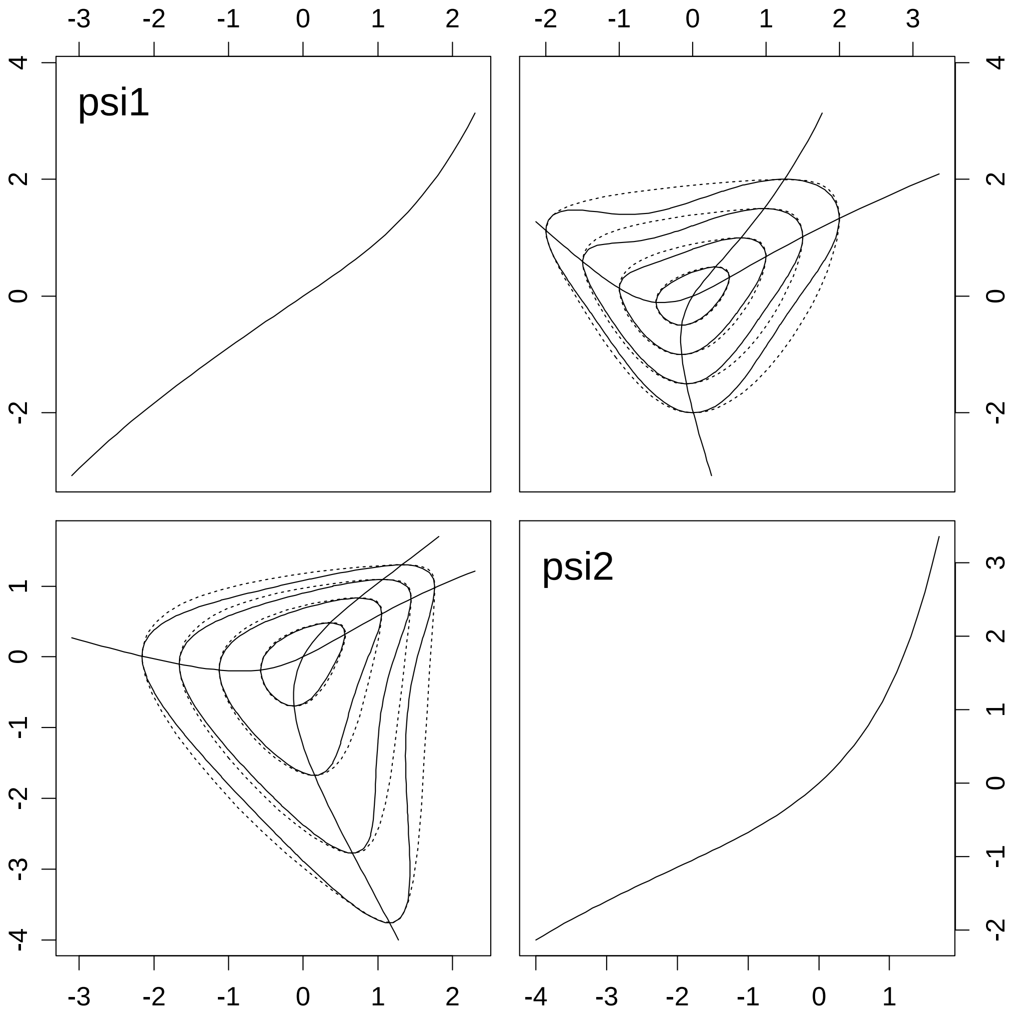
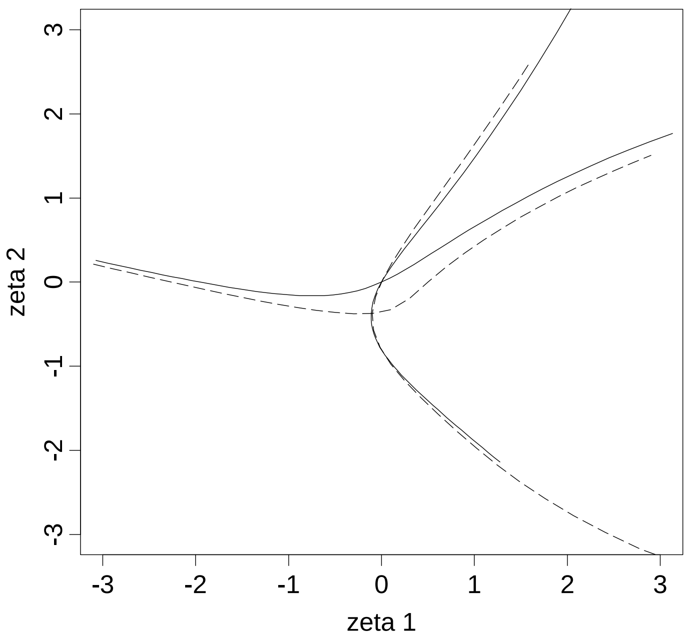


Figure 6: Profile pairs diagnostic plot for ϕ coordinates. Solid lines: Profiles and exact contours computed by evaluation of F over fine grids. Dotted lines: Approximate contours obtained with the interpolation method of Bates and Watts.



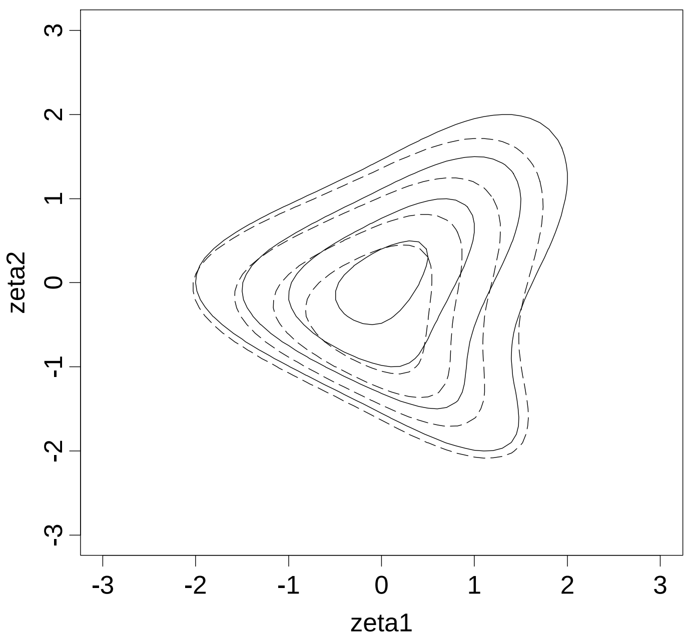


Figure 7: Top: Transformed profile traces (solid) versus profile traces in the transformed coordinates (dashed) for the Example 2. Bottom: Contours of the transformed log-posterior (solid) versus contours of the posterior in the transformed coordinates (dashed).

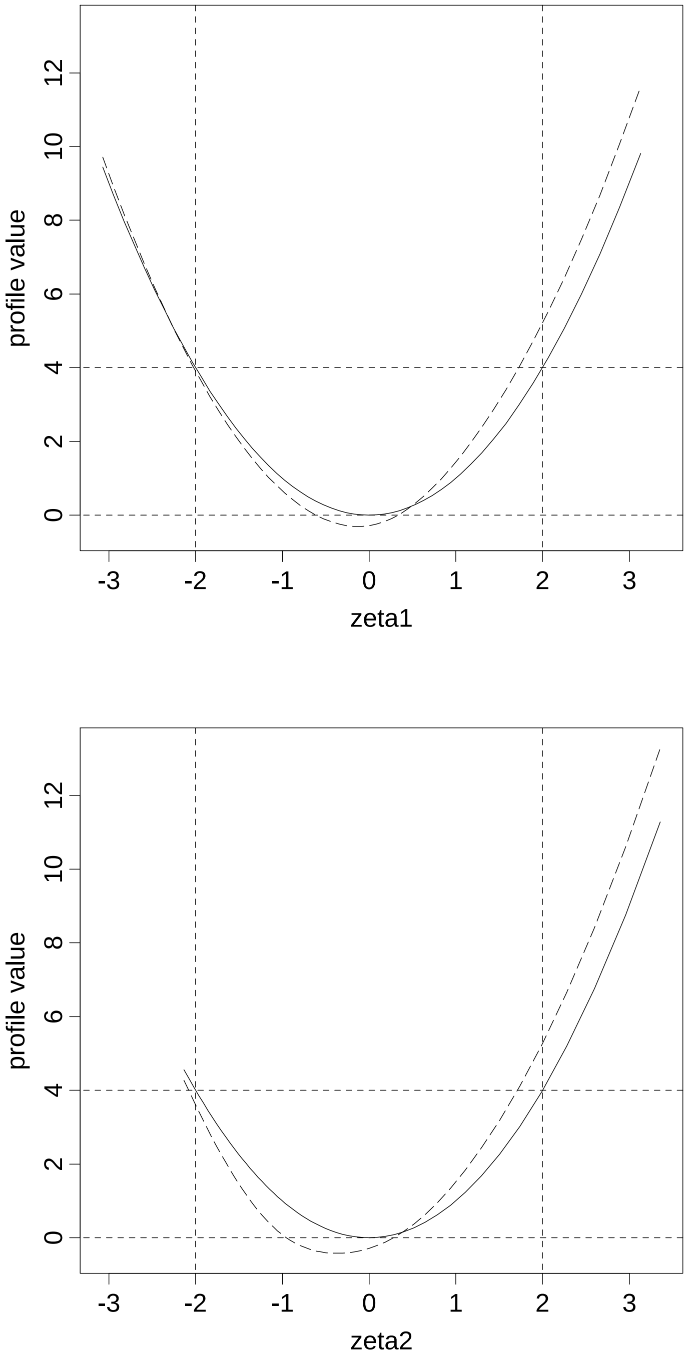


Figure 8: Top: First profile value on transformed log posterior (solid) versus first profile value after considering the Jacobian of the transformation (dashed). Bottom: Second profile value on transformed log posterior (solid) versus second profile value after considering the Jacobian of the transformation (dashed).

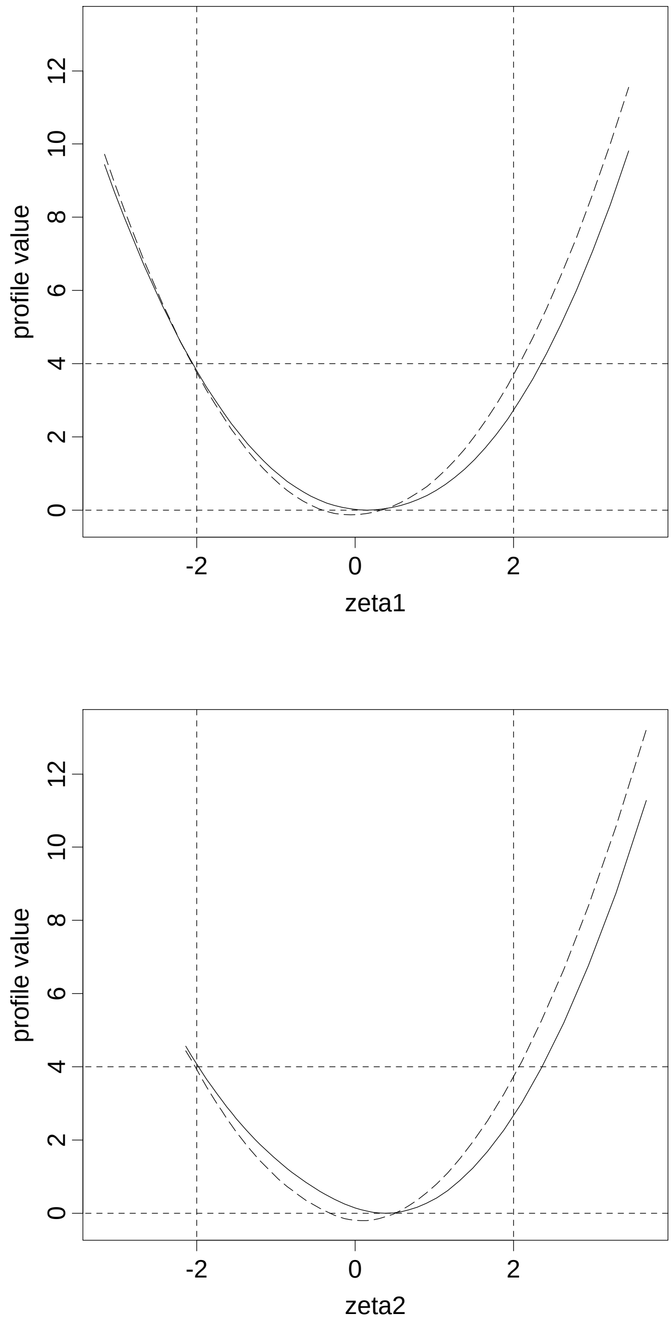


Figure 9: Profile values after one iteration of the adjustment algorithm. Top: First profile value on transformed log posterior (solid) versus first profile value after considering the Jacobian of the transformation (dashed). Bottom: Second profile value on transformed log posterior (solid) versus second profile value after considering the Jacobian of the transformation (dashed).