Design for nonlinear mixed-effects Are variances a reasonable scale?

Douglas Bates

University of Wisconsin - Madison
<Bates@Wisc.edu>

PODE, Paris, France
March 22, 2012

Outline

Outline

Outline

Outline

Outline

Outline

Outline

D-optimal experimental design for fixed-effects models

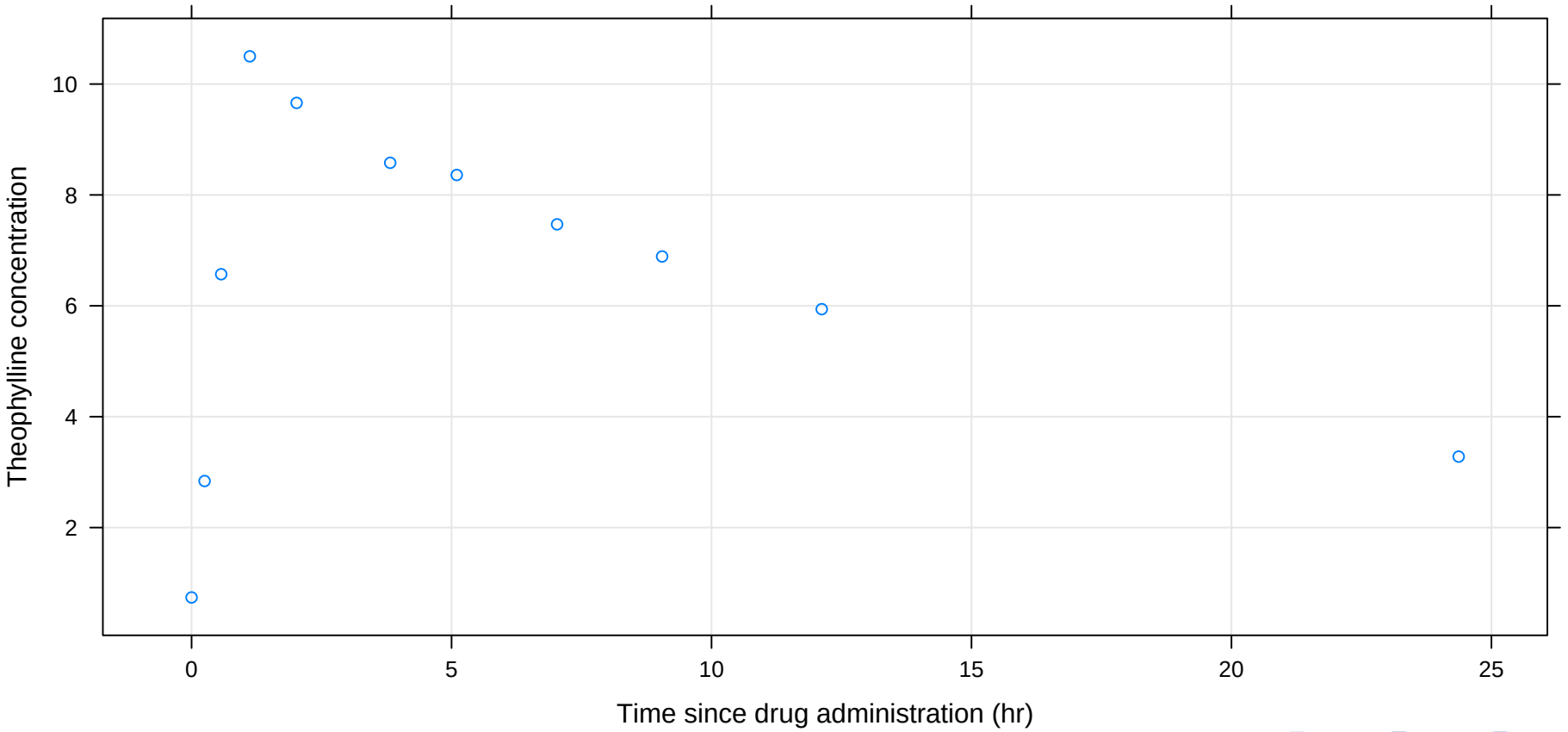
- The purpose of D-optimal experimental design is to minimize the volume of confidence regions or likelihood contours or HPD regions on the parameters.
- For simple cases (e.g. linear models with no random effects) the choice of parameters does not affect the design. In some ways the only parameters that make sense are the coefficients of the linear predictor and these are all equivalent up to linear transformation.
- For a nonlinear model the choice of parameters is less obvious. Nonlinear transformations of parameters can produce dramatically better or worse linear approximations. In terms of likelihood contours or H.P.D. regions the ideal shape is elliptical (i.e. a locally quadratic deviance function) but the actual shape can be quite different.

D-optimal design for mixed-effects models

- For a linear mixed-effects model the choice of scale of the variance components affects the shape of deviance or posterior density contours.
- For a nonlinear mixed-effects model, both the scale of the variance components and the choice of model parameters affect the shape of such contours.
- These distorsions of shape are more dramatic when there are fewer observations per group (i.e. per `Subject` or whatever is the grouping factor). But that is exactly the situation we are trying to achieve.

Profiling nonlinear regression models

- This is a very brief example of profiling nonlinear regression models with a change of parameters.
- Take data from a single subject in the `Theoph` data set in `R`



My initial naive fit

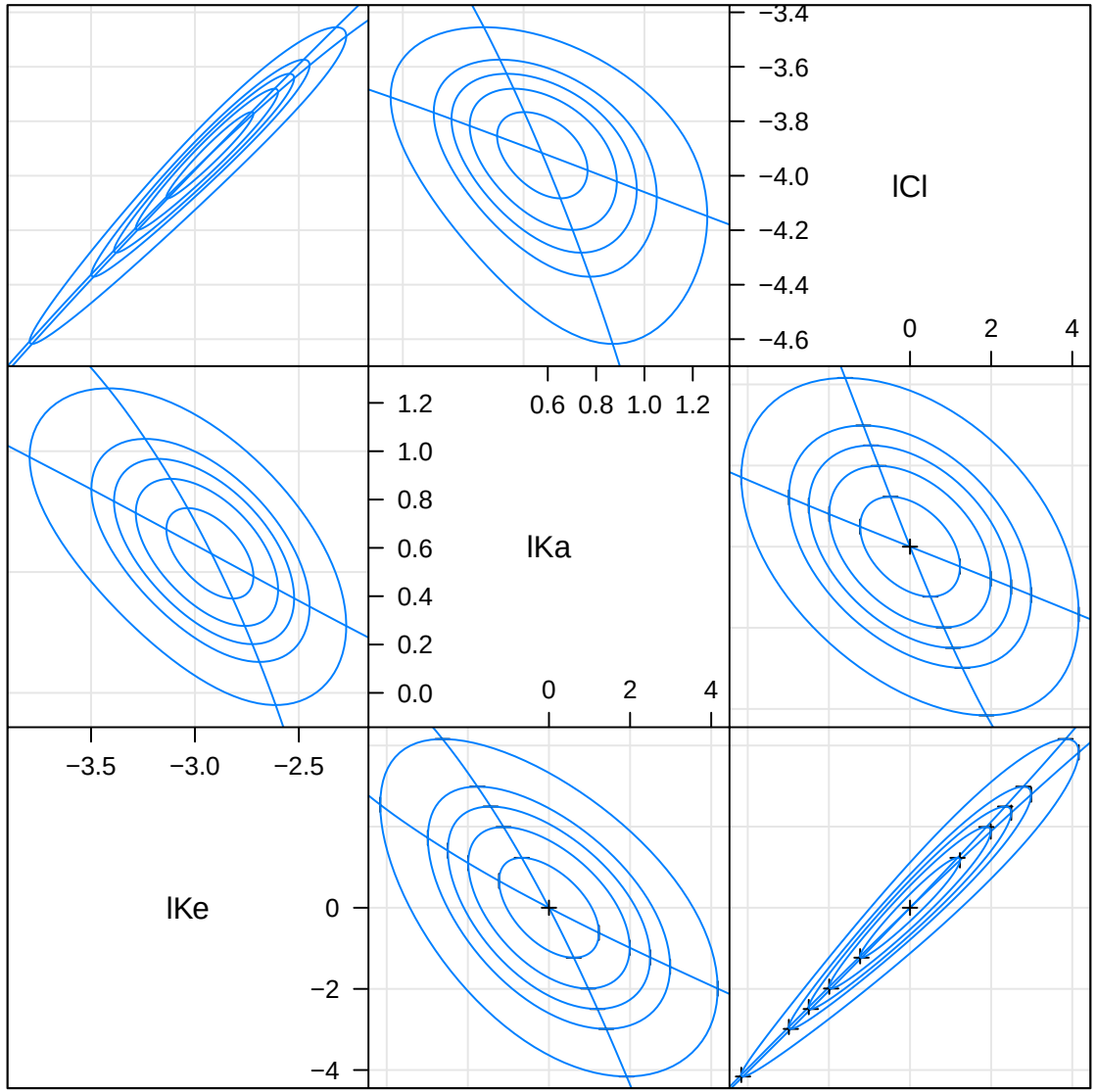
```
> Theo.1 <- droplevels(subset(Theoph, Subject==1))
> summary(fm1 <- nls(conc ~ SSfol(Dose, Time, lKe, lKa, lCl), T
```

```
Formula: conc ~ SSfol(Dose, Time, lKe, lKa, lCl)
Parameters:
    Estimate Std. Error t value Pr(>|t|)
lKe  -2.9196     0.1709 -17.085 1.40e-07
lKa   0.5752     0.1728   3.328   0.0104
lCl  -3.9159     0.1273 -30.768 1.35e-09

Residual standard error: 0.732 on 8 degrees of freedom

Correlation of Parameter Estimates:
    lKe   lKa
lKa -0.56
lCl  0.96 -0.43

Number of iterations to convergence: 8
Achieved convergence tolerance: 4.907e-06
```

Following a suggestion from France Mentré

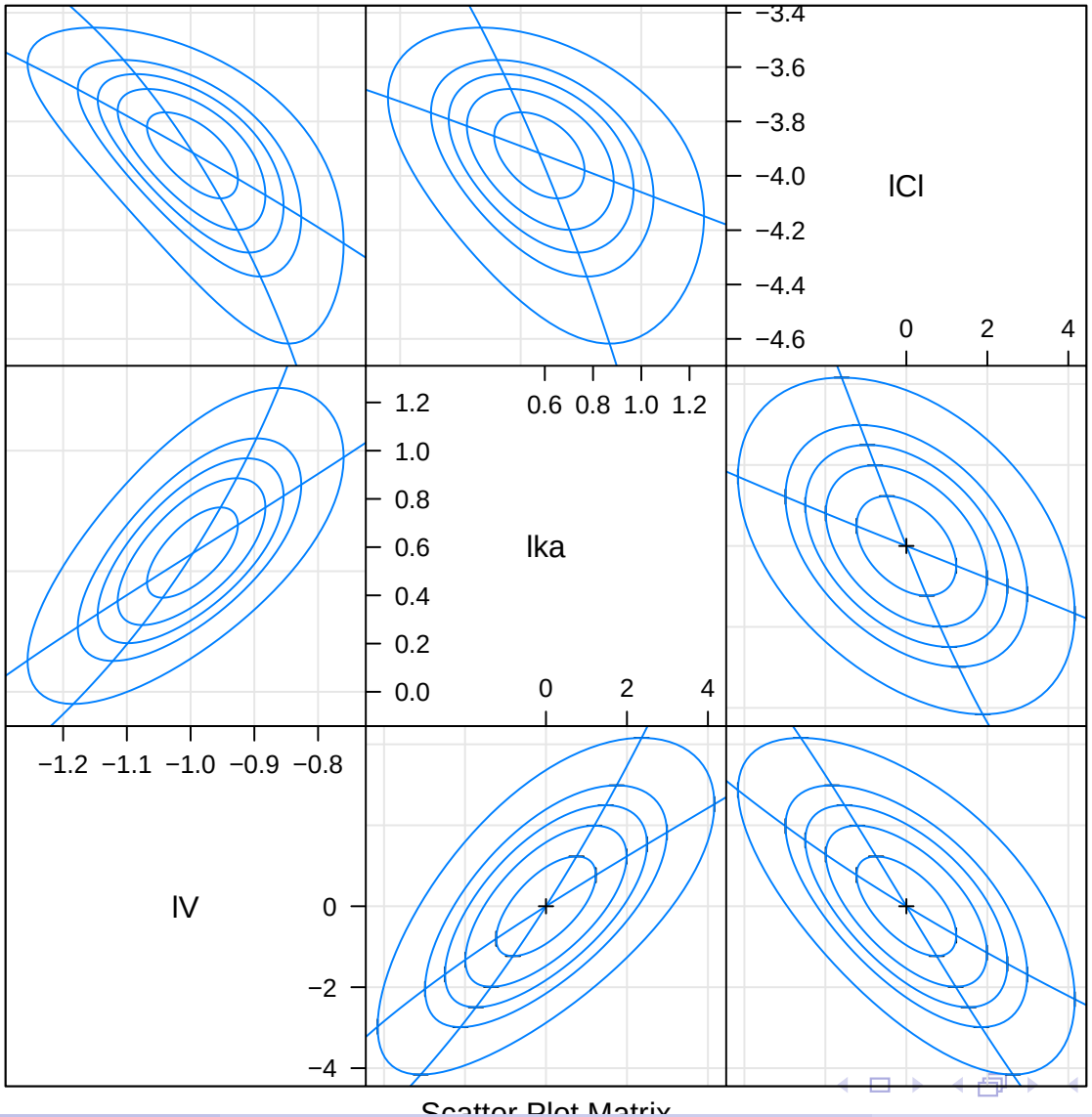
```
> oral1cptSdlkalVlCl <- PKmod("oral", "sd", list(ka ~ exp(lka)
> summary(gm1 <- nls(conc ~ oral1cptSdlkalVlCl(Dose, Time, lV,

Formula: conc ~ oral1cptSdlkalVlCl(Dose, Time, lV, lka, lCl)
Parameters:
    Estimate Std. Error t value Pr(>|t|)
lV  -0.99624    0.06022 -16.543 1.80e-07
lka  0.57516    0.17282   3.328   0.0104
lCl -3.91586    0.12727 -30.768 1.35e-09

Residual standard error: 0.732 on 8 degrees of freedom

Correlation of Parameter Estimates:
    lV    lka
lka  0.68
lCl -0.61 -0.43

Number of iterations to convergence: 9
Achieved convergence tolerance: 4.684e-06
```

Contours based on profiling the objective, original



Scatter Plot Matrix

Contours based on profiling the objective, revised formulation



Estimates based on optimizing a criterion

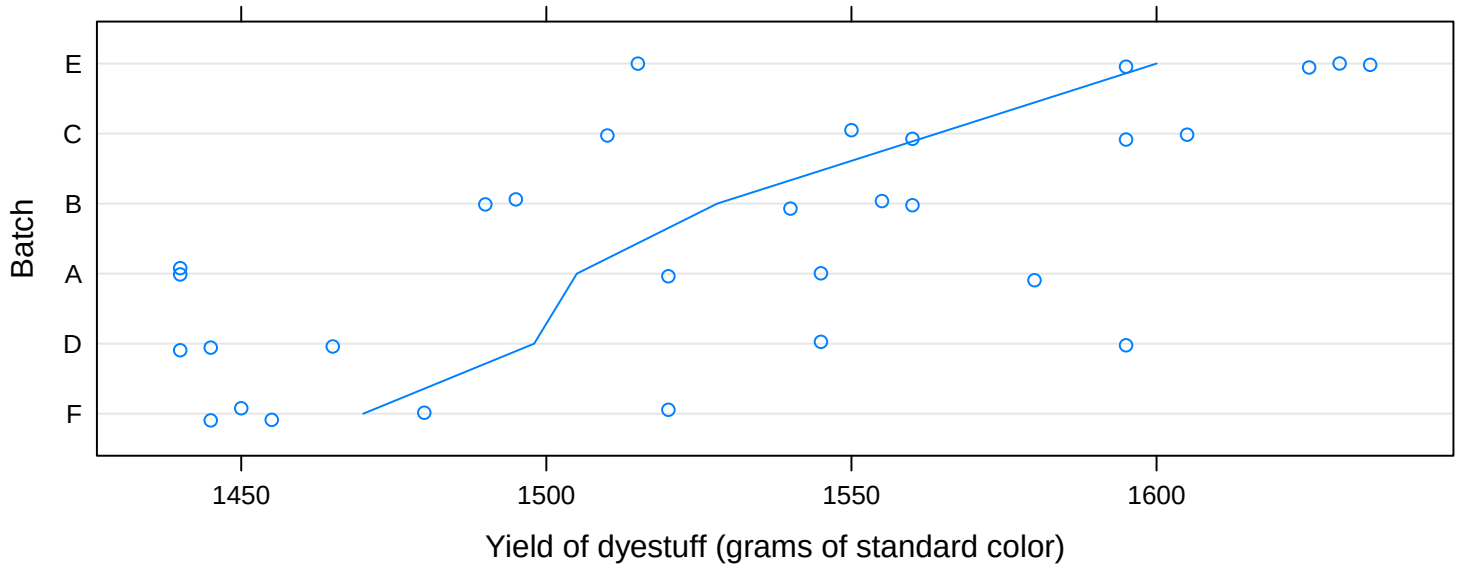
- Maximum-likelihood estimators are an example of estimators defined as the values that optimize a criterion – maximizing the log-likelihood or, equivalently, minimizing the deviance (negative twice the log-likelihood).
- Deriving the distribution of such an estimator can be difficult (which is why we fall back on asymptotic properties) but, for a given data set and model, we can assess the sensitivity of the objective (e.g. the deviance) to the values of the parameters.
- We can do this systematically by evaluating one-dimensional "profiles" of the objective, through conditional optimization of the objective.

Profiling the objective

- Profiling is based on conditional optimization of the objective, fixing one or more parameters at particular values and optimizing over the rest.
- We will concentrate on one-dimensional profiles of the deviance for mixed-effects models but the technique can be used more generally.
- We write the deviance as $d(\boldsymbol{\phi}|\boldsymbol{y})$ where $\boldsymbol{\phi}$ is the parameter vector of length p and $\boldsymbol{y}$ is the vector of observed responses. Write the individual components of $\boldsymbol{\phi}$ as $\phi_k, k = 1, \ldots, p$ and the complement of ϕ_i as $\boldsymbol{\phi}_{-i}$.
- The profile deviance is $\tilde{d}_i(\phi_i) = \min_{\boldsymbol{\phi}_{-i}} d((\phi_i, \boldsymbol{\phi}_{-i})|\boldsymbol{y})$. The values of the other parameters at the optimum form the *profile traces*
- If estimates and standard errors are an adequate summary then the deviance should be a quadratic function of $\boldsymbol{\phi}$, i.e. $\tilde{d}_i(\phi_i)$ should be a quadratic centered at $\hat{\phi}_i$ and the profile traces should be straight.

A Simple Example: the Dyestuff data

The `Dyestuff` data in the `lme4` package for `R` are from the the classic book *Statistical Methods in Research and Production*, edited by O.L. Davies and first published in 1947.



The line joins the mean yields of the six batches, which have been reordered by increasing mean yield.

The effect of the batches

- The particular batches observed are just a selection of the possible batches and are entirely used up during the course of the experiment.
- It is not particularly important to estimate and compare yields from these batches. Instead we wish to estimate the variability in yields due to batch-to-batch variability.
- The `Batch` factor will be used in *random-effects* terms in models that we fit.
- In the "subscript fest" notation such a model is

$$y_{i,j} = \mu + b_i + \epsilon_{i.j}, \quad i = 1, \ldots, 6;\ j = 1, \ldots, 5$$

  with $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$ and $b_i \sim \mathcal{N}(0, \sigma_1^2)$.
- We obtain the maximum-likelihood estimates for such a model using `lmer` with the optional argument, `REML=FALSE`.

Fitted model

```
> (fm1 <- lmer(Yield ~ 1 + (1|Batch), Dyestuff, REML=FALSE))

Linear mixed model fit by maximum likelihood ['lmerMod']
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff
      AIC      BIC   logLik deviance
 333.3271 337.5307 -163.6635 327.3271

Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 1388     37.26
 Residual             2451     49.51
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1527.50      17.69   86.33
```

Profiling the fitted model

```
> head(pr1 <- profile(fm1))
```

```
      .zeta    .sig01   .sigma (Intercept)   .par
1 -2.3243980  0.000000 61.96437      1527.5 .sig01
2 -2.2270119  6.315107 60.74307      1527.5 .sig01
3 -1.9527642 12.317030 57.71369      1527.5 .sig01
4 -1.5986116 17.365631 54.69985      1527.5 .sig01
5 -1.1872929 22.297821 52.32154      1527.5 .sig01
6 -0.7326184 27.624184 50.72564      1527.5 .sig01
```

```
...
```

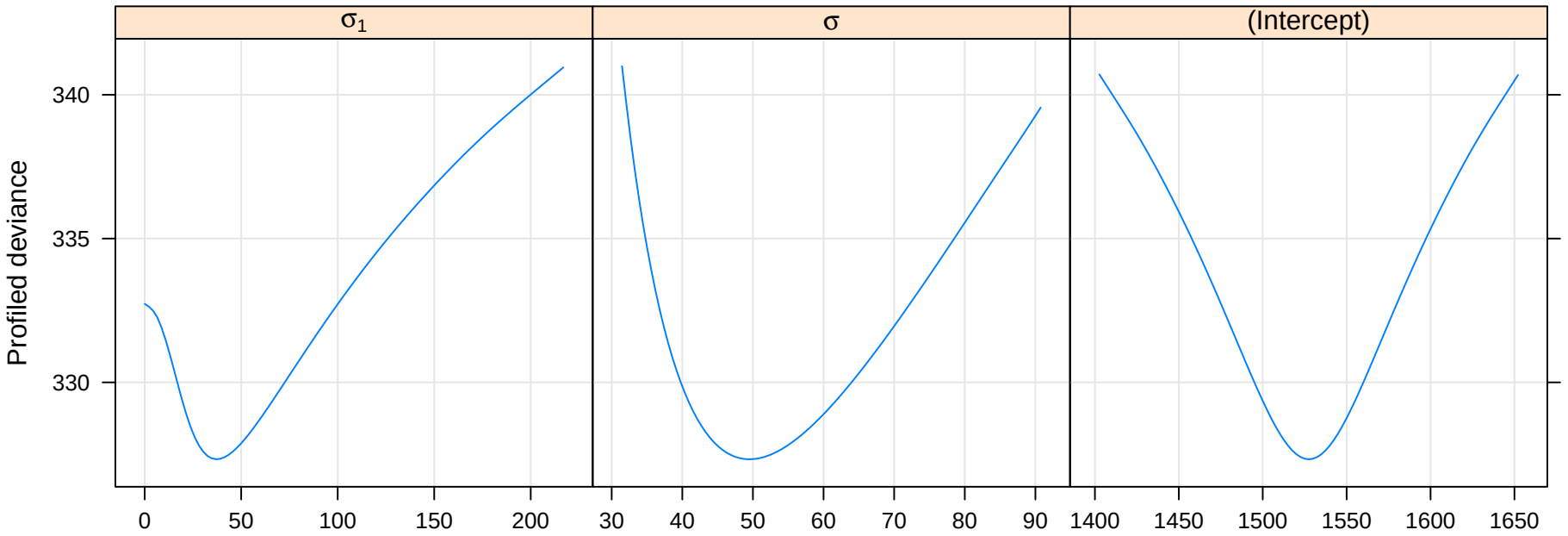
```
      .zeta    .sig01  .sigma (Intercept)        .par
55 1.950491  55.23929 49.5101    1568.280 (Intercept)
56 2.305893  63.77264 49.5101    1579.255 (Intercept)
57 2.653119  74.71123 49.5101    1592.257 (Intercept)
58 2.993207  88.72455 49.5101    1608.022 (Intercept)
59 3.327036 106.75187 49.5101    1627.538 (Intercept)
60 3.655342 130.10234 49.5101    1652.153 (Intercept)
```

Reconstructing the profiled deviance

In `pr1` the profiled deviance, $\tilde{d}_i(\phi_i)$ is expressed on the *signed square root* scale

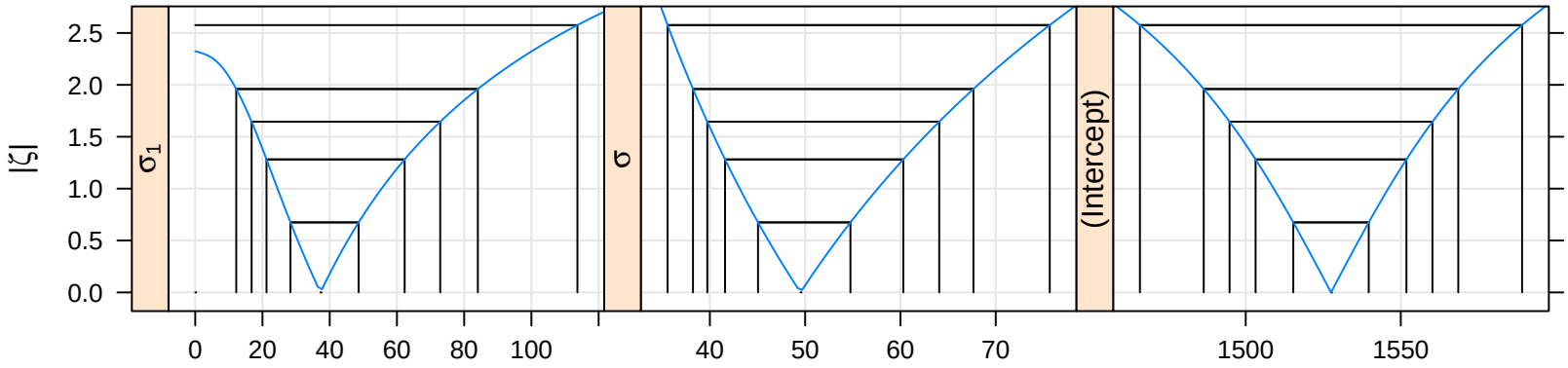
$$\zeta_i(\phi_i) = \operatorname{sgn}(\phi_i - \hat{\phi}_i)\sqrt{\tilde{d}_i(\phi_i) - d(\widehat{\boldsymbol{\phi}}|\boldsymbol{y})}$$

In the original scale, $\tilde{d}_i(\phi_i)$, the profiles are



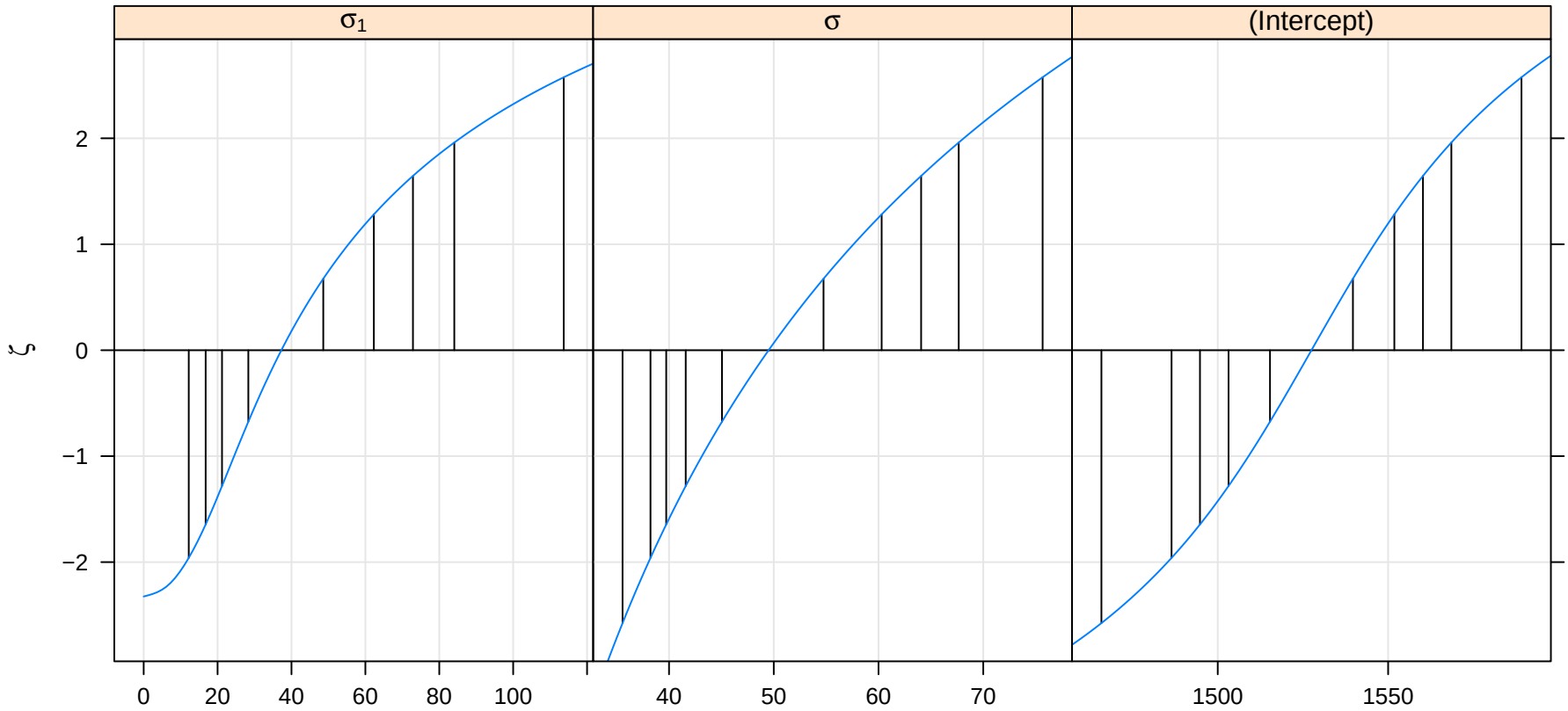
After applying the square root

On the scale of $\sqrt{\tilde{d}_i(\phi_i) - d(\widehat{\boldsymbol{\phi}}|\boldsymbol{y})}$ the profiles are



We have added intervals corresponding to 50%, 80%, 90%, 95% and 99% confidence intervals derived from the profiled deviance.

And, finally, on the ζ scale

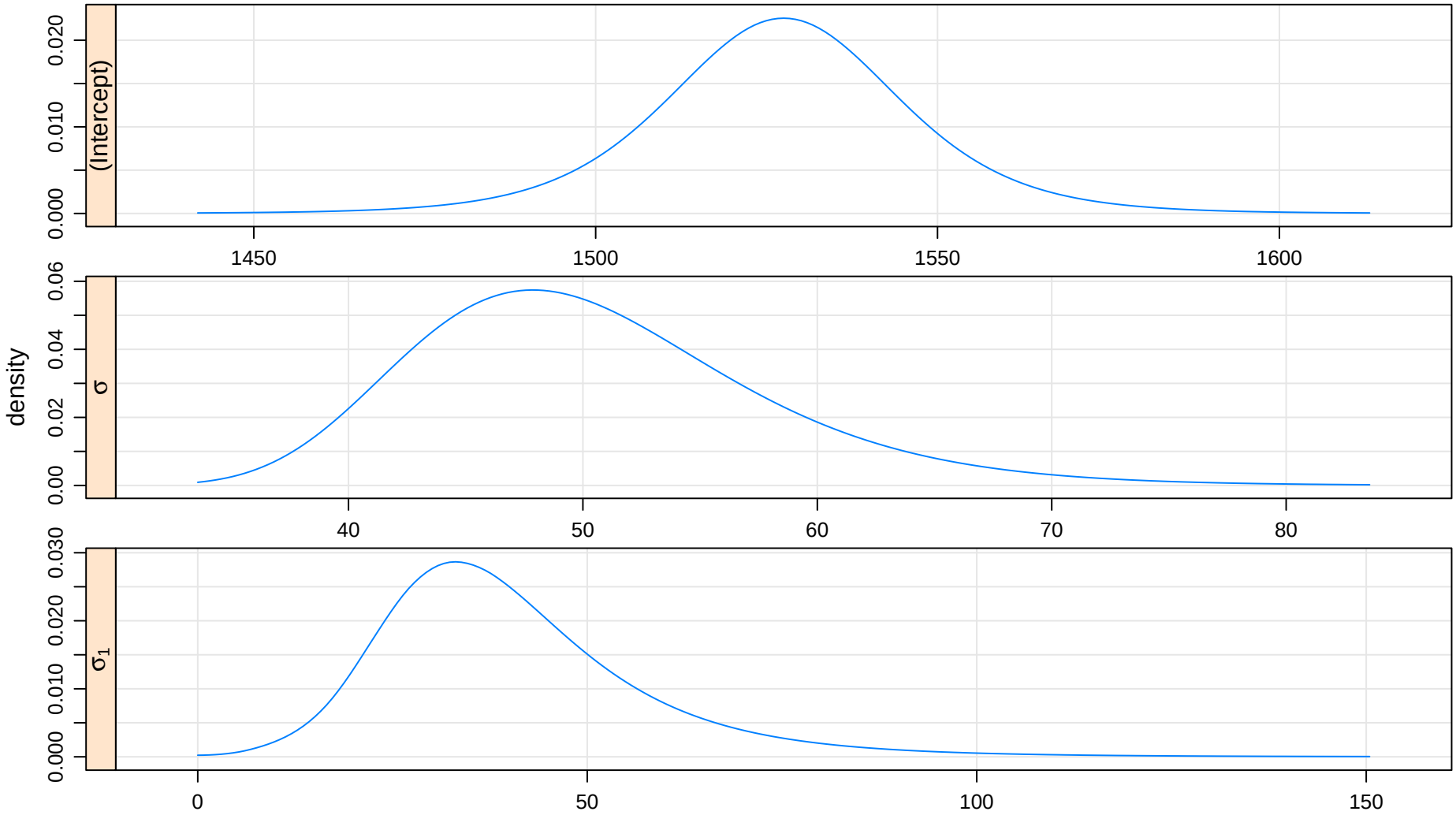


The intervals are created by "inverting" likelihood ratio tests on particular values of the parameter.

Representing profiles as densities

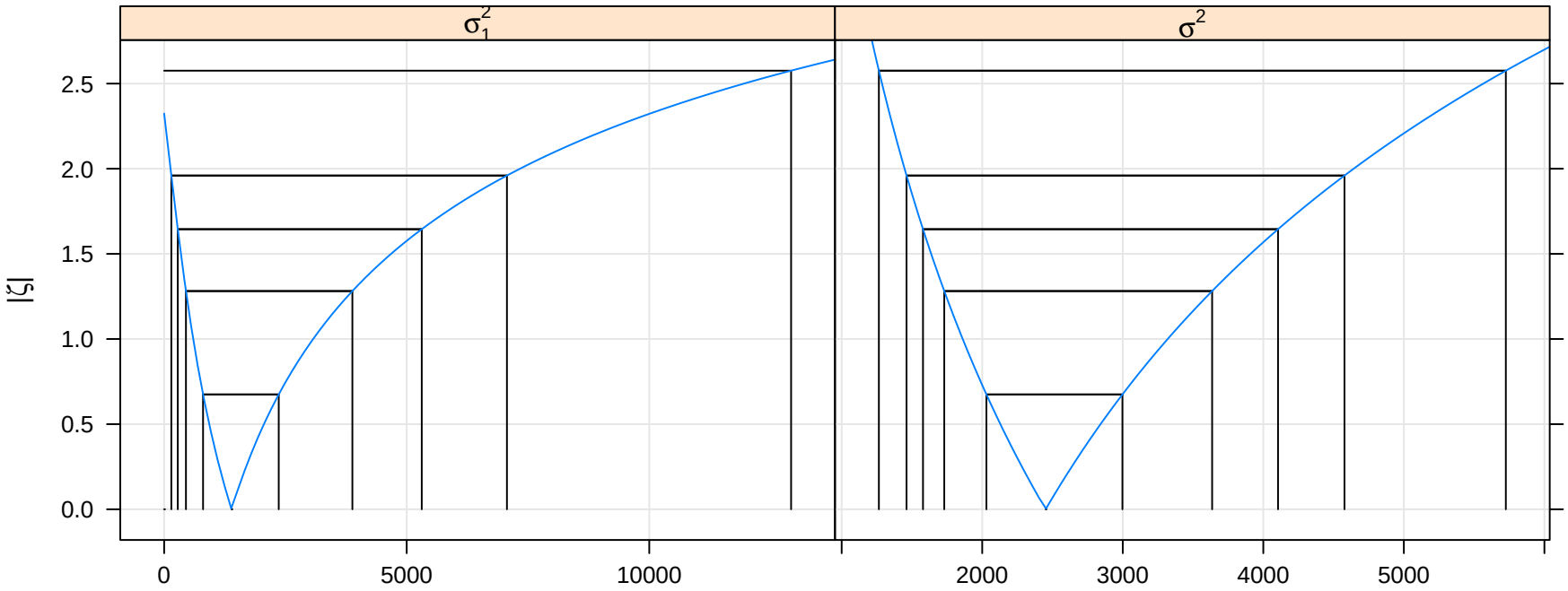
- A univariate profile ζ plot is read like a normal probability plot
  - a sigmoidal (elongated "S"-shaped) pattern like that for the `(Intercept)` parameter indicates overdispersion relative to the normal distribution.
  - a bending pattern, usually flattening to the right of the estimate, indicates skewness of the estimator and warns us that the confidence intervals will be asymmetric
  - a straight line indicates that the confidence intervals based on the quantiles of the standard normal distribution are suitable
- If we map the ζ scale through the cdf, Φ, for the standard normal, $\mathcal{N}(0,1)$, we can derive a cdf and a density for a distribution that would produce this profile.

Profiles for parameters in `fm1` as densities

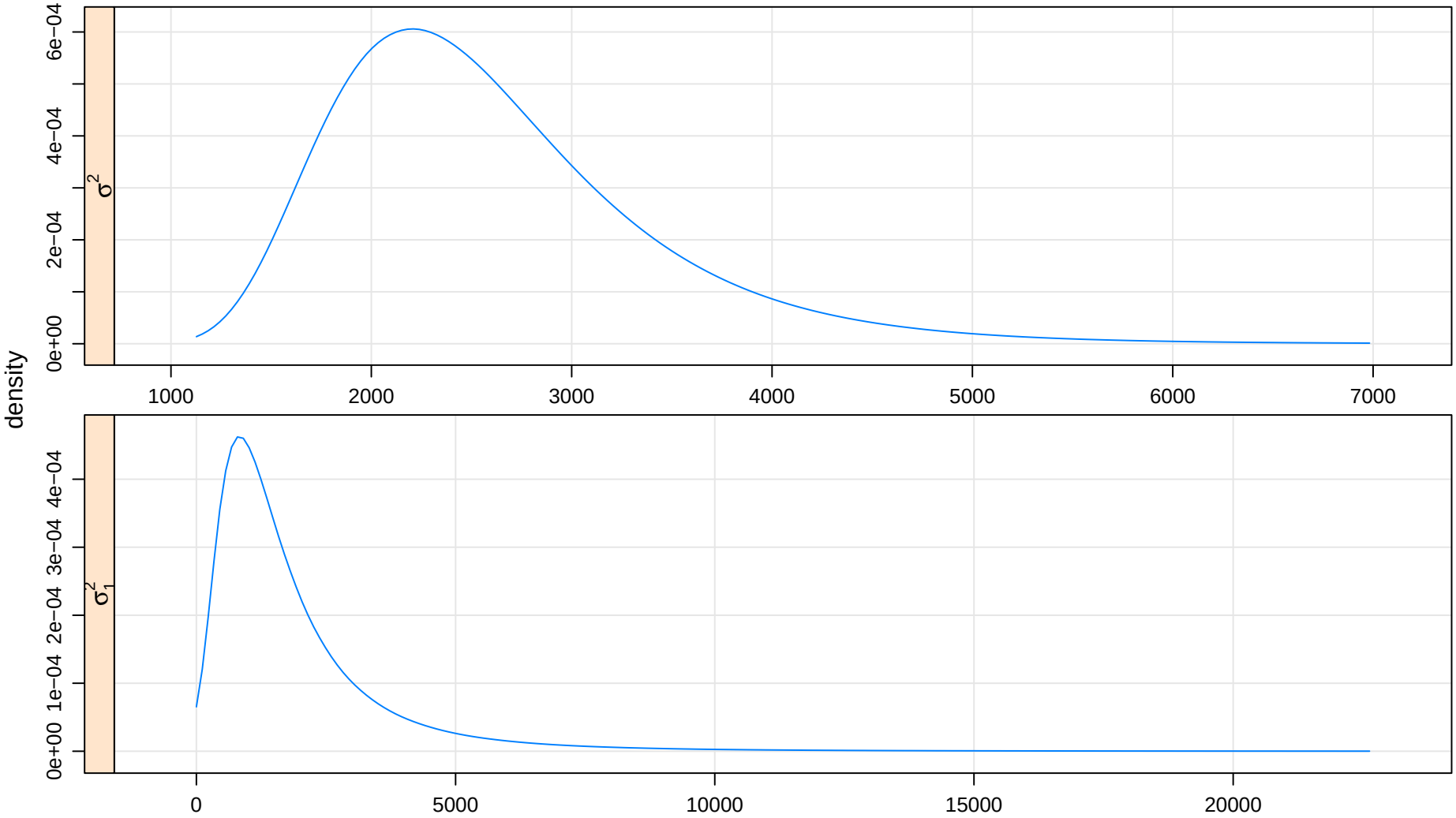


Profile ζ plots on the scale of the variance components

Usually the variability estimates in a mixed-effects model are quoted on the scale of the "variance components", σ_1^2 and σ^2, not the standard deviations as we have shown. On the variance scale the profiles are



Densities of variance components

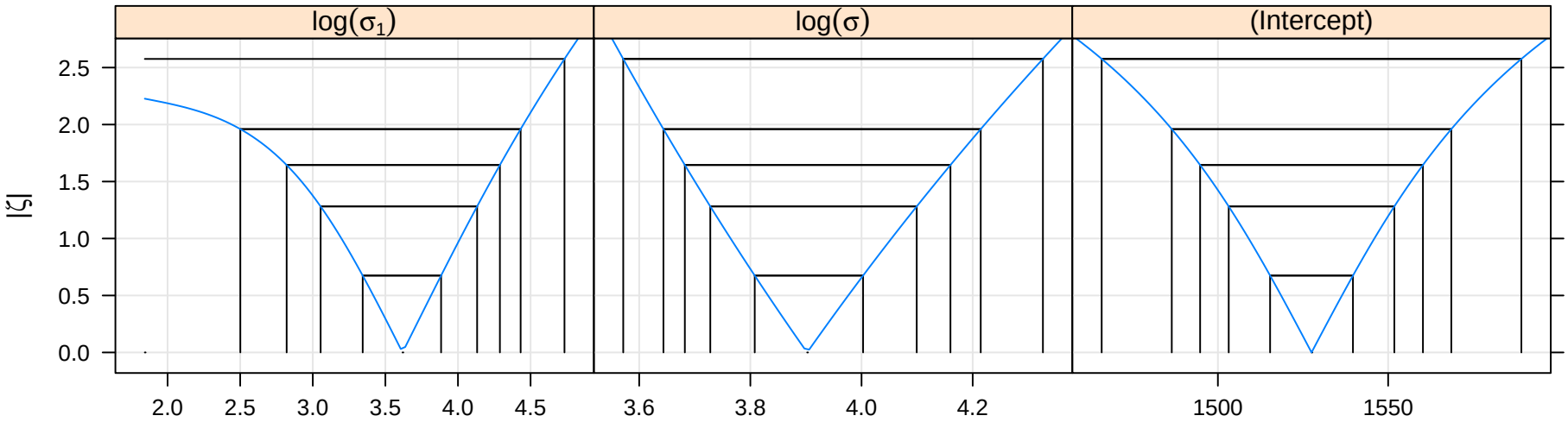


Some practical implications

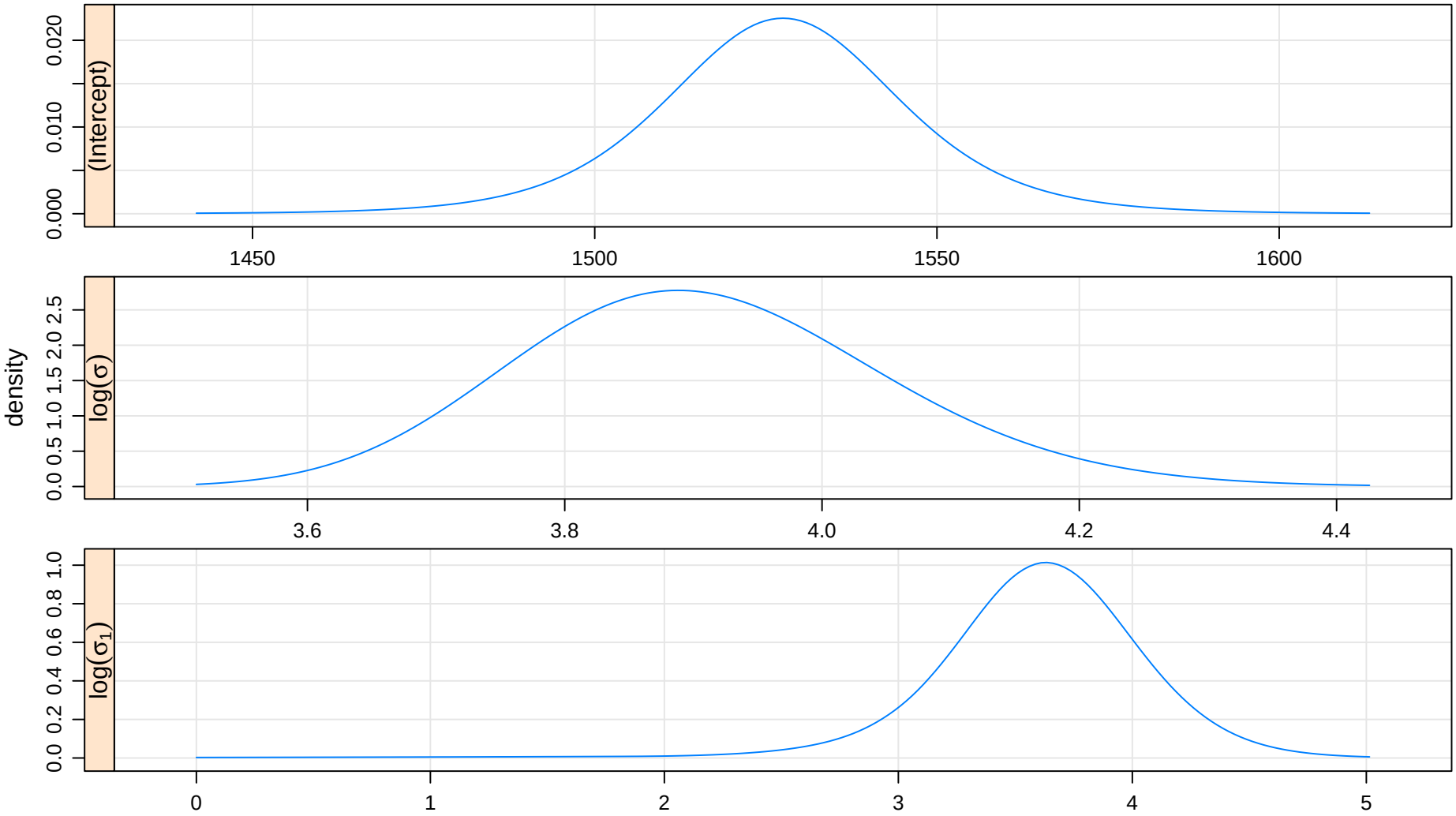
- We have been using maximum likelihood estimates. For linear mixed-effects models the REML estimates are often preferred because they are assumed to be less biased. (Many people assume they are unbiased but, except in certain special cases, they're not.)
- But bias is a property of the expected value of the estimator. So bias of a variance estimator relates to the mean of one of those badly skewed distributions. Why should we use the mean?
- Similarly, it is common in simulation studies to compare estimators or computational methods based on mean squared error. That's not a meaningful criterion for skewed distributions of estimators.

A more reasonable scale

Distributions of the estimators are closer to being symmetric on the scale of $\log(\sigma)$ and $\log(\sigma_1)$ (or, equivalently, $\log(\sigma^2)$ and $\log(\sigma_1^2)$) except when 0 is in the range of reasonable values,



Densities on the logarithm scale

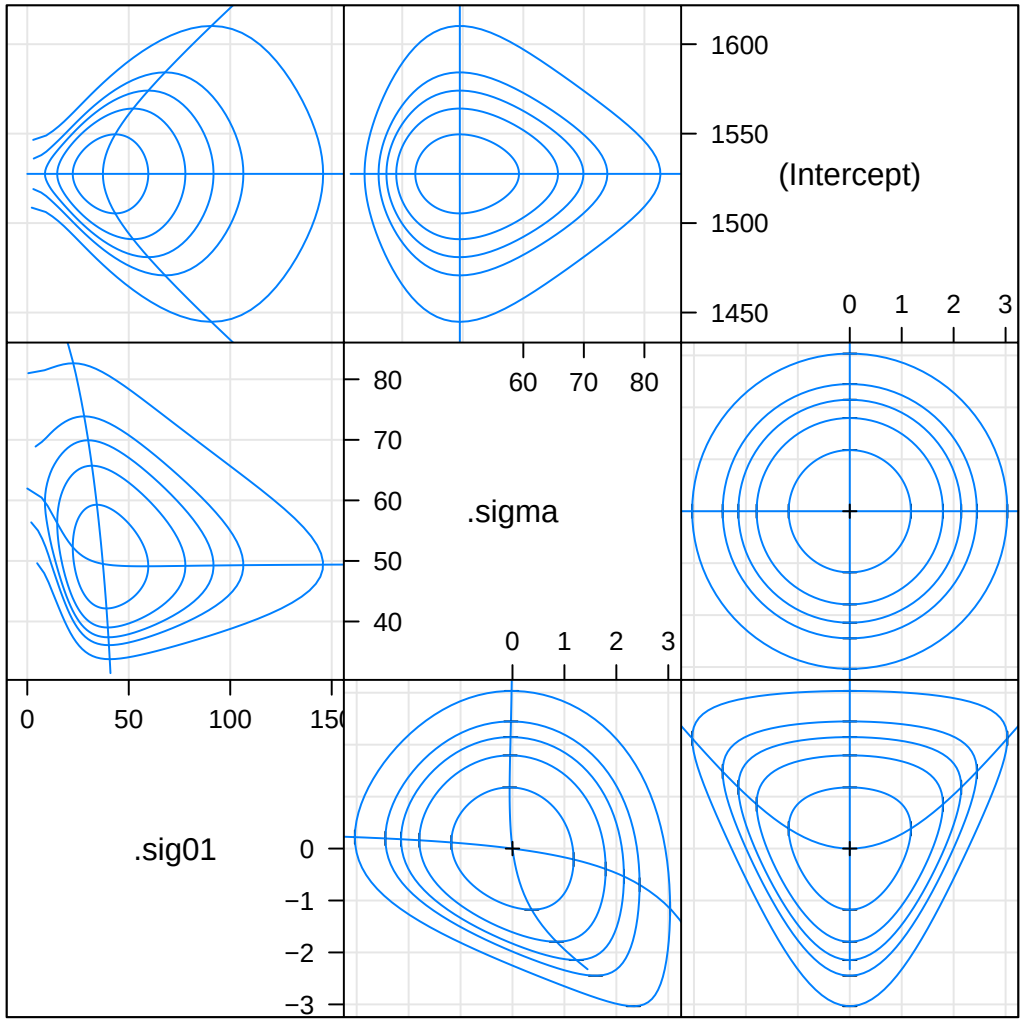


Profile pairs plots

- The information from the profile can be used to produce pairwise projections of likelihood contours. These correspond to pairwise joint confidence regions.
- Such a plot (next slide) can be somewhat confusing at first glance.
- Concentrate initially on the panels above the diagonal where the axes are the parameters in the scale shown in the diagonal panels. The contours correspond to 50%, 80%, 90%, 95% and 99% pairwise confidence regions.
- The two lines in each panel are "profile traces", which are the conditional estimate of one parameter given a value of the other.
- The actual interpolation of the contours is performed on the ζ scale which is shown in the panels below the diagonal.

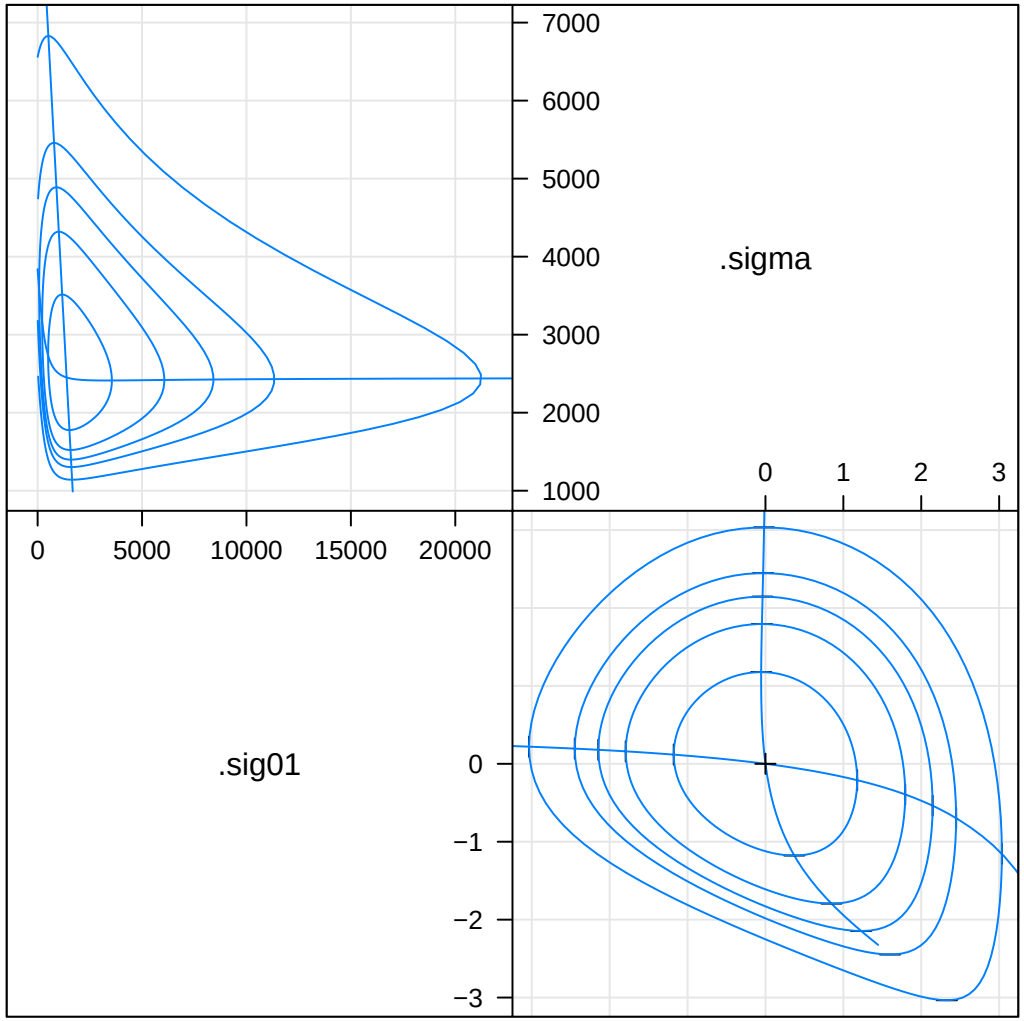
Profile pairs for model

```
> splom(pr1)
```



Scatter Plot Matrix

Profile pairs for the variance components



Scatter Plot Matrix