Faster Data Fitting Solver for Calibration

Novel Nonlinear Least Squares Trust Region Solver from the NAG Library

Solver Introduction

In the NAG Library (Mark 27.1 onwards) we provide a novel nonlinear least squares (NLN-LSQ) trust-region solver handle_solve_bxnl (`e04gg`) for unconstrained and bound-constrained fitting problems which implements various algorithms and regularization techniques. It is aimed at small to medium sized fitting problems (up to 1000s of parameters) of the form

$$\min_{x \in \mathbb{R}^{n_{\text{var}}}} \quad \frac{1}{2}\sum_{i=1}^{n_{\text{res}}}[w_i r_i(x)]^2 + \frac{\sigma}{p}\|x\|_2^p$$
$$\text{subject to} \quad l_x \le x \le u_x$$

where $r_i(x), i = 1, \ldots, n_{\text{res}}$, are smooth nonlinear functions called residuals, $w_i, i = 1, \ldots, n_{\text{res}}$ are weights (by default they are all defined to 1, and the rightmost element represents the regularization term with parameter $\sigma \ge 0$ and power $p > 0$. The constraint elements l_x and u_x are n_{var}-dimensional vectors defining the bounds on the variables.

Calibrating the parameters of complex numerical models to fit real-world observations is one of the most common problems found in industries such as finance, physics, simulations, engineering, etc. Typically in a calibration or data fitting context, the residuals will be defined as the difference between the observed values y_i at t_i and the values provided by a nonlinear model $\phi(t;x)$, i.e.,

$$r_i(x) := y_i - \phi(t_i; x).$$

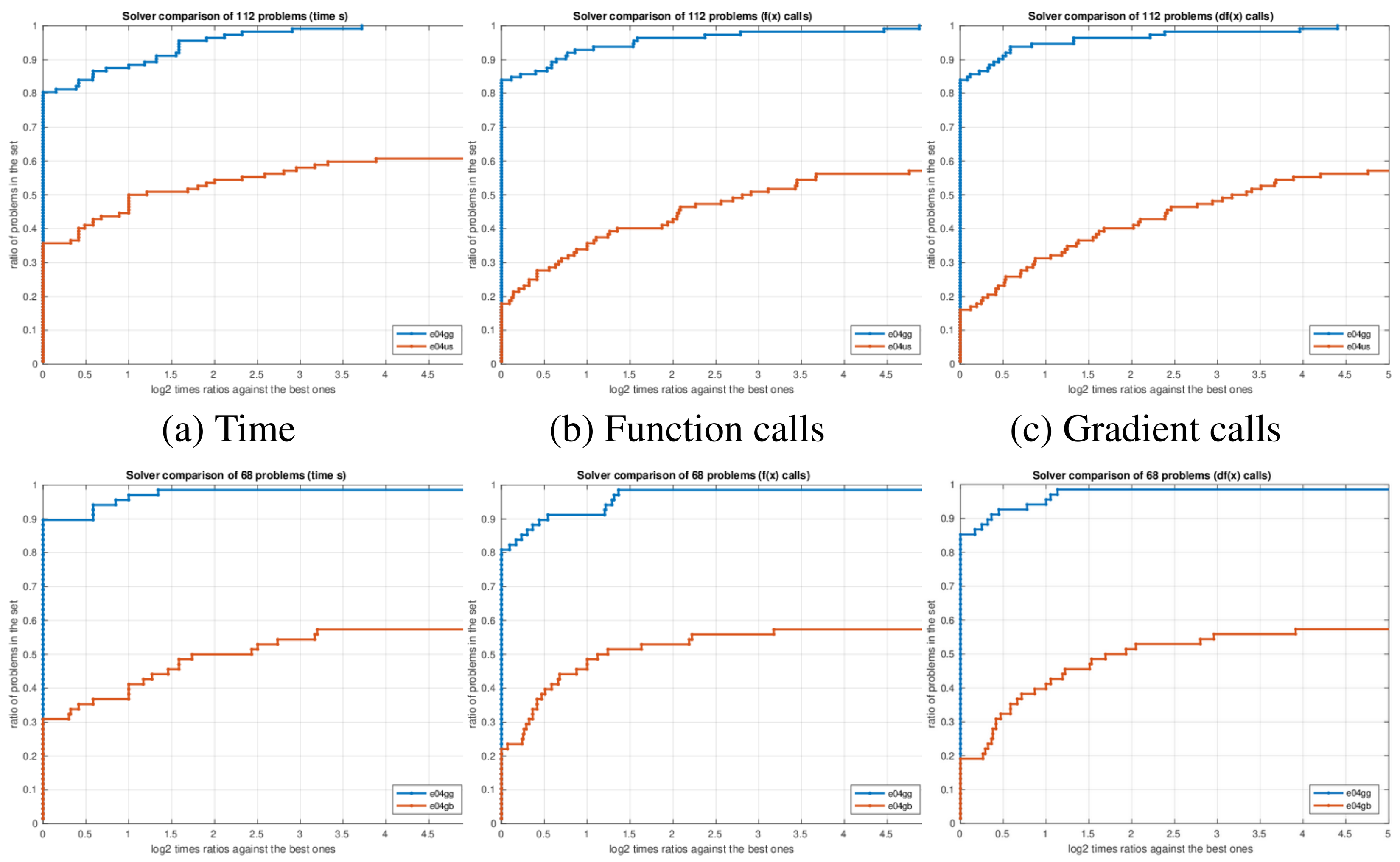
The new solver `e04gg` provides a significant improvement over the current nonlinear least squares solvers in the NAG Library. In addition, this solver fills the gap between unconstrained solvers like lsq_uncon_quasi_deriv_comp (`e04gb`) and the fully constrained ones, like lsq_gencon_deriv (`e04us`).

Features of `e04gg`

- Well established trust-region method with a **variety of implemented algorithms**, ranging from simple **Powell's dogleg** or **Gauss-Newton methods** to sophisticated **Tensor–Newton schemes** which overcome convergence difficulties present in simpler methods;
- **Avoids data over-fitting** by incorporating different types of regularization for both the problem formulation and the trust-region subproblem;
- **Ability to recover** when it is not possible to evaluate the function $\phi(t, x)$ or its gradient at a given point;
- **Account for uncertainty** in the observed data by using optional residual weights;
- **Flexible stopping criteria** that can suit the problem and data tolerances.

Modern Replacement for `e04gb` and `e04us`

The new solver `e04gg` offers unprecedented robustness and a significant speed-up over current alternatives in the Library, namely `e04gb` for unconstrained nonlinear least squares problems and `e04us` for problems with simple variable bounds. You are highly recommended to upgrade to the new solver.



(a) Time (b) Function calls (c) Gradient calls

(d) Time (e) Function calls (f) Gradient calls

Figure 1: Performance profiles

Our benchmarks comparing `e04gg` to `e04gb` on 68 unconstrained nonlinear least squares CUTEst problems are shown in Figure 1 (a)–(c) using performance profiles. The three plots show `e04gg` is: faster, solving 60% of the problems in less time (a); more robust, solving 25% more problems; and requires fewer user callbacks, with 55% of problems requiring fewer function calls (b) and 65% requiring fewer gradient evaluations (c).

As the `e04gb` solver was not designed to tackle simple bounds directly, `e04us` can be used instead for such problems. However, `e04us` being a more general solver does not fully exploit the structure of NLN-LSQ problems as `e04gg` does in the presence of simple bounds. That's why a speed-up on 45% of problems can be observed (d) as well as a reduction in user callbacks: 65% of problems require fewer function and gradient calls (e and f) when comparing `e04us` and `e04gg` on 112 unconstrained and bound-constrained nonlinear least square CUTEst problems.

Real-world: α Particle Track Data Fitting

This example illustrates the use of `e04gg` to fit PADC etched nuclear track data to a convoluted distribution. Track diameteres from a scanned target sheet (Figure 2) are recorded into a histogram (blue bars in Figure 3). Then we fit a mixed Normal and log-Normal model to the experimental histogram.

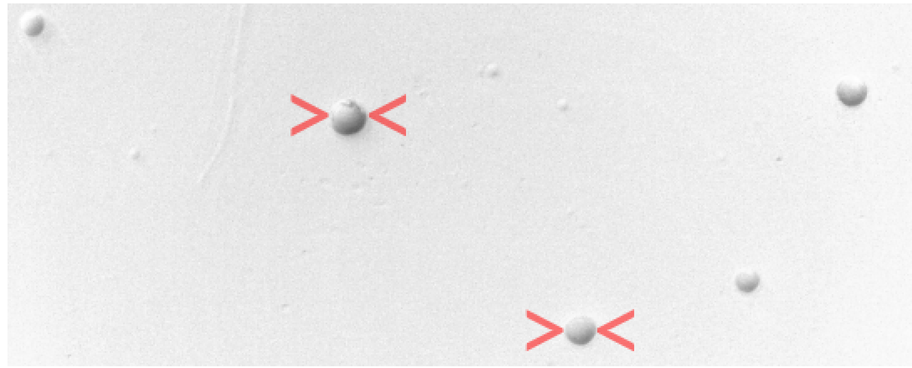


Figure 2: α particle etched tracks

The `e04gg` solver is used to fit the six parameter model

$$\phi\big(t, x = (a, b, A_\ell, \mu, \sigma, A_g)\big) = \text{log-Normal}(a, b, A_l) + \text{Normal}(\mu, \sigma^2, A_g)$$
$$\text{subject to} \quad 0 \le x,$$

using the histogram heights.

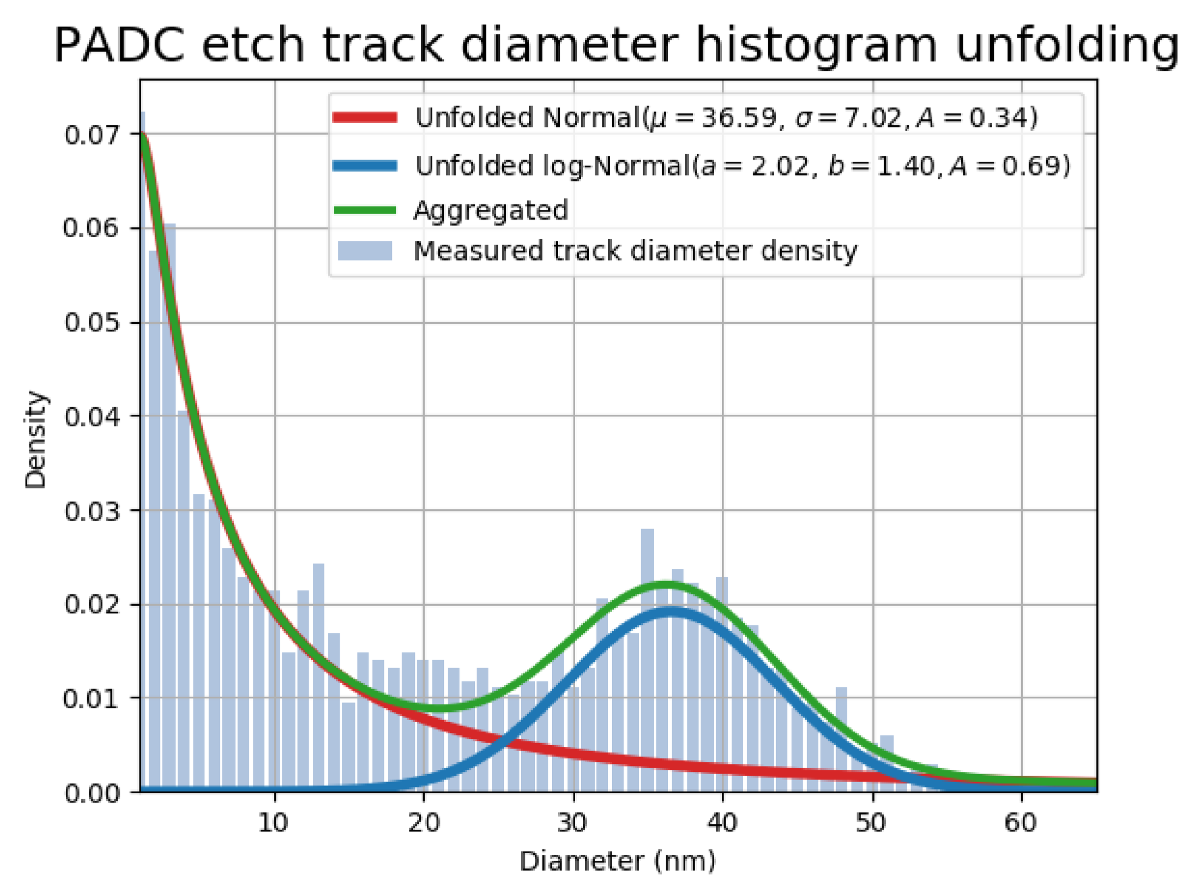


Figure 3: Aggregated model fitted to diameter histogram and unfolded models. Optimal parameter values shown in legend

Thanks to the use of regularization and residual weights, `e04gg` provided a robust solution x^* to unfold the parameters for the two distributions (red and blue curves in Figure 3). Adding these together produce the green curve which is the one used to perform the fitting.

Find more examples *with* source code at: `github.com/numericalalgorithmsgroup`

support@nag.com *www.nag.com*