

Applications and Extensions of a Technique for Estimator Densities

Peter Hingley *

Abstract—Applications are given of a formula for the exact probability density function of the maximum likelihood estimates of a statistical model, where the data generating model is allowed to differ from the estimation model. The main examples are supported by simulation experiments. Curved exponential families are investigated, for which an approach is described that can be used in many practical situations. The distribution of a maximum likelihood estimator in exponential regression is developed. Non-linear regression is then considered, with an example of a model discrepancy situation arising in ELISA immunoassays and similar biochemical titrations. An incorrect logistic model is specified for a titration curve that is used for describing the reaction of a chemical sample to applied substrate concentration. A method is suggested to reduce the amount of bias in the estimate of binding affinity. Finally there is a prospective discussion of other possible uses of the technique, including general comparisons of sets of alternative models in frequentist and Bayesian settings, applications to robust estimation and extensions beyond maximum likelihood estimates.

Keywords: biochemical titration, ELISA, exponential regression, M-estimators, maximum likelihood, robustness, nonlinear regression, technique for estimator densities

1 Introduction

A technique for estimator densities (TED) gives the exact joint density of the maximum likelihood estimates (MLE) from a specified statistical model, typically a nonlinear regression model [11]. The method can be used where the estimation model either agrees with or differs from the model that has generated the data.

The use of a specific estimation model is widespread when the data are presumed to be distributed in a certain way according to a scientific hypothesis. Nevertheless the modeller may accept that alternative hypotheses are possible. TED considers a pair of models without exploring specifically the question of discriminating between them. The models are freely chosen and need not be nested. We can consider two situations.

1. Estimation model *equivalent* to data generating model. Here TED is in competition with existing approximate and exact analytic techniques. It is an addition to the statistical toolbox as an analytical approach to derive the exact algebraic expression for the density of the MLE.

2. Estimation model *not equivalent* to data generating model. Here TED may be the only exact analytic method that is available for describing the density of the quasi maximum likelihood estimates (which will also be termed MLE here).

Since TED operates under both of these situations, it can be used as a basis for assessing the robustness of an estimation model against deviations from the presumed data generating process. An exact criterion can also be constructed that is based on Kullback-Leibler information for the comparison of a pair of alternative models as fitted to a set of data. This dispenses with the asymptotic approximation that is inherent in most other such criteria in common use like AIC [12]. TED is of most potential value for cases where data samples are unique or expensive to replicate. It can therefore be expected to be particularly useful in areas such as epidemiology (e.g. [16]) and econometrics (e.g. [5]), where bias can arise from functional differences between models or by overfitting or underfitting models to data. It should be stressed that the construction of the analytic density of the MLE is an algebraic exercise that can be intractable for more complicated setups. TED facilitates the algebraic procedure but may still not provide a closed solution in more difficult cases.

In this paper the application of TED to curved exponential families will be described and two examples will be given. The first example is exponential regression, where the technique gives an easy path to derive results that are already available in the literature. The second example demonstrates nonlinear regression modelling in the setting of biochemical titration experiments, where distinct data generating model and estimation model are specified. For these examples S-Plus has been used because of its easy facility for handling vectors and matrices [25]. APL2 was also previously found to be a suitable medium for the calculations [10][11]. Finally there is a discussion of ways that TED could be extended to other kinds of estimates, including in particular M-estimates for ro-

*European Patent Office, Landsberger Strasse 187, D-80687 Munich, Germany. Email: phingley@epo.org

bust estimation, and also how comparison of members of a family of models can be approached via TED under either frequentist or Bayesian paradigms.

2 TED applied to curved exponential family estimation models

2.1 The technique

In the following, statistical models will be specified in terms of the densities of data that are generated by them.

The n members of a sample are described as a $(n \times 1)$ vector w , $g_0(w)$ is the true density of w , and $g_1(w|\theta)$ is the presumed density with a $(p \times 1)$ parameter vector θ to be estimated. The log likelihood corresponding to $g_1(w|\theta)$ is $l(\theta|w)$. The space of w is W , and the space of θ is Θ . A $(n \times 1)$ vector of independent variables z or a design matrix can be introduced to cope with the regression situation.

$\hat{\theta}$ is the MLE and is given as $l'(\theta, w)|_{\theta=\hat{\theta}} = 0$, (where $'$ indicates differentiation with respect to θ).

Consider a $(p \times 1)$ vector T .

$$T(\theta, \theta^*, w) = l'(\theta^*, w) - l'(\theta, w), \tag{1}$$

where θ^* is fixed at an arbitrary value. Under a simple set of regularity conditions, the exact density for $\hat{\theta}$ is given as follows.

$$g(\hat{\theta}) = E_w[[j(\theta, w)|_{\theta=\hat{\theta}}] \cdot g_{[T(\hat{\theta}, \theta^*=\hat{\theta}, w)]}(0), \tag{2}$$

where $j(\theta, w) = -l''(\theta, w)$ is the observed information, and the second term represents the value of the density $g_{[T(\hat{\theta}, \theta^*, w)]}(t)$, for which $\theta^* = \hat{\theta}$, and hence $t = 0$ by (1). The term $E_w[[j(\theta, w)|_{\theta=\hat{\theta}}]$ describes a conditional expectation, that is conditional on $\theta = \hat{\theta}$ and is taken with respect to w over $g_1(w|\theta)$.

$$E_w[[j(\theta, w)|_{\theta=\hat{\theta}}] =$$

$$\left[\frac{\int_{\widehat{\theta(v)}} |j(\theta, w(v))|_{\theta=\hat{\theta}} \cdot g_1(w(v)|\theta=\hat{\theta}) \cdot ||w'(v)|| dv}{\int_{\widehat{\theta(v)}} g_1(w(v)|\theta=\hat{\theta}) \cdot ||w'(v)|| dv} \right] \tag{3}$$

It is usually unnecessary to evaluate the multidimensional integrals in (3), because in practice terms in w can be replaced by $E_w[w|\theta = \hat{\theta}]$, terms in w^2 by $E_w[w^2|\theta = \hat{\theta}]$, etc. For example, if the model $g_1(w|\theta)$ was normal $N(h[\theta_1], \theta_2)$, terms proportional to w would be replaced by terms proportional to $h[\hat{\theta}_1]$, and terms proportional to w^2 would be replaced by terms proportional to $\hat{\theta}_2 + [h(\hat{\theta}_1)]^2$.

The proof of (2) is given in [11] and consists of three parts. Part 1 considers the special case where each realisable value of $\hat{\theta}$ is associated with a distinct data vector,

$w_{\hat{\theta}}$. The conditional density of $\hat{\theta}$ for fixed θ^* , $h(\hat{\theta}|\theta^*)$ say, is then given by a standard change of variable argument, using a Jacobian. Part 2 of the proof goes on to consider the usual case that more than one $w_{\hat{\theta}}$ vector can exist for each $\hat{\theta}$. Now the conditional density of $\hat{\theta}$, $g(\hat{\theta}|\theta^*)$ say, is given as a conditional expectation of $h(\hat{\theta}|\theta^*)$, conditional on $\hat{\theta}$ over the range of $w_{\hat{\theta}}$ vectors. Part 3 of the proof provides the unconditional density $g(\hat{\theta})$, by taking account of the fact that, for each $\hat{\theta}$ value, part 2 specifies a distinct density $g(\hat{\theta}|\theta^* = \hat{\theta})$, for which $T(\hat{\theta}, \theta^* = \hat{\theta}, w(v)) = 0$, by equation (1) and the property $l'(\theta, w)|_{\theta=\hat{\theta}} = 0$. The description of the density $g(\hat{\theta})$ is made for various distinct $\hat{\theta}$ values and, for each of them, $\theta^* = \hat{\theta}$ can be selected. When taken together, the corresponding terms $g(\hat{\theta}|\theta^* = \hat{\theta})$ constitute a new density $g(\hat{\theta})$ that is independent of θ^* and is given by (2).

A simple introductory example involving the MLE of the mean of a normal distribution estimation model is given in [12]. Further examples are considered below.

2.2 TED with curved exponential family estimation models

TED will now be considered in the context of curved exponential families because a straightforward formulation is obtained and many useful estimation models are covered.

Following Dobson [4], let θ appear in a $(n \times 1)$ canonical function $b(\theta, z)$ and in a $(p \times 1)$ functional $c(\theta, z)$, together with $(n \times 1)$ functionals of the data $a(w)$ and $d(w)$; all constrained to describe a valid density for w .

$$g_1(w|\theta) = \exp[a(w)^T b(\theta, z) + 1_{(p \times 1)}^T c(\theta, z) + 1_{(n \times 1)}^T d(w)], \tag{4}$$

where T indicates transposition.

If $f(\theta, z)$ is the unconditional expectation of $a(w)$, then the $(p \times 1)$ vectors for $l'(\theta, w)$ and $T(\hat{\theta}, \theta^*, w)$ are given as follows.

$$l'(\theta, w) = b'(\theta, z) \cdot (a(w) - f(\theta, z)) \tag{5}$$

$$T(\hat{\theta}, \theta^*, w) = b'(\theta^*, z) \cdot (a(w) - f(\theta^*, z)) \tag{6}$$

Equation (6) shows that $T(\hat{\theta}, \theta^*, w)$ is a linear transform of $a(w)$, and so $g_{[T(\hat{\theta}, \theta^*, w)]}(t)$ can often be found easily.

The conditional expectation $E_w[[j(\theta, w)|_{\theta=\hat{\theta}}]$ is obtained as in (3) from the observed information $j(\theta, w)$, which is a $(p \times p)$ matrix that is calculated by differentiating $l'(\theta, w)$ again.

$$-j(\theta, w) = l''(\theta, w) =$$

$$[b''(\theta, z) \cdot (a(w) - f(\theta, z))] - [f'(\theta, z) \cdot (b'(\theta, z))^T] \tag{7}$$

Here $b''(\theta, z)$ is a $(p \times p \times n)$ matrix, while $f'(\theta, z)$ and $b'(\theta, z)$ are $(p \times n)$ matrices.

3 Exponential regression (Case 1, estimation model equivalent to data generating model)

The prescriptive formula (2) will often be easier to use than other suggested analytic methods. Consider for example an equation for $g(\hat{\theta})$ that was introduced by Hillier and Armstrong [7], and then applied to exponential regression by Hillier and O'Brien [8]. Here, equivalents to a subset of their results will be demonstrated using TED, which offers a shorter derivation.

Data are distributed according to a negative exponential density, with the rate parameter itself given as an exponential function of an underlying independent ($n \times 1$) variate z . As above, the data set is of n values w_i with prespecified independent variables z_i . A single scalar parameter θ_0 is to be estimated in this example ($p = 1$).

$$g_0(w_i) = \gamma_{0i} \cdot \exp[-w_i \gamma_{0i}],$$

$$w_i > 0, \quad \gamma_{0i} = \exp[-z_i \theta_0] > 0 \quad (8)$$

This can be written in vector forms for the equivalent data generating model and estimation models.

$$g_0(w) = \exp[-\theta_0 z^T \mathbf{1} - w^T \cdot \exp[-\theta_0 z_i]] \quad (9)$$

$$g_1(w|\theta) = \exp[-\theta z^T \mathbf{1} - w^T \exp[-\theta z_i]] \quad (10)$$

Here, $\mathbf{1}$ is a ($n \times 1$) vector of 1s. $\exp[-\theta_0 z_i]$ and $\exp[-\theta z_i]$ are ($n \times 1$) vectors that contain, respectively, the scalar quantities $\exp[-z_i \theta_0]$ and $\exp[-z_i \theta]$ taken over the n values of z_i ($i = 1, \dots, n$).

Density (10) is a member of the exponential family (4), with

$$a(w) = w, \quad b(\theta, z) = -\exp[-\theta z_i],$$

$$1^T c(\theta, z) = -\theta z^T \mathbf{1}, \quad 1^T d(w) = 0.$$

The unconditional expectation of $a(w)$ is the ($n \times 1$) vector $f(\theta, z) = \exp[\theta z_i]$. So, from (6),

$$T(\hat{\theta}, \theta^*, w) = (z_i \exp[-\theta^* z_i])^T \cdot (w - \exp[\theta^* z_i]), \quad (11)$$

where $z_i \exp[-\theta^* z_i]$ and $\exp[\theta^* z_i]$ are ($n \times 1$) vectors.

Also

$$-j(\theta, w) = l''(\theta, w) = -(z_i^2 \exp[-\theta z_i])^T \cdot w$$

$$E_w[[j(\theta, w)|\theta=\hat{\theta}] = E_w[(z_i^2 \exp[-\theta z_i])^T \cdot w|_{\theta=\hat{\theta}}] =$$

$$[(z_i^2 \exp[-\hat{\theta} z_i])^T] \cdot (E_w[w|_{\theta=\hat{\theta}}] =$$

$$(z_i^2 \exp[-\hat{\theta} z_i])^T \cdot (\exp[\hat{\theta} z_i]) = (z_i^2)^T \cdot \mathbf{1}, \quad (12)$$

where $z_i^2 \exp[-\theta z_i]$ and z_i^2 are ($n \times 1$) vectors.

Examining (11), $T(\hat{\theta}, \theta^*, w)$ is a weighted sum of exponentials with an offset. Let

$$V = (z_i \exp[-\theta^* z_i])^T \cdot w = \sum_{i=1}^n w_i z_i \cdot \exp[-z_i \theta^*] \quad (13)$$

Then

$$T(\hat{\theta}, \theta^*, w) = V - (z_i \exp[-\theta^* z_i])^T \cdot (\exp[\theta^* z_i]) = V - z^T \mathbf{1} \quad (14)$$

The density of T can be obtained by first finding the density of V and then applying a transformation. If v_i is a standard exponential variable ($g(v_i) = \exp[-v_i]$), then V can be expressed as a weighted sum of n such independent variables. From (13), $V = \sum_{i=1}^n \Phi_i v_i$, where, in terms of scalar quantities,

$$\Phi_i = z_i \exp[-z_i \theta^*] \cdot \exp[z_i \theta_0] = z_i \exp[z_i(\theta_0 - \theta^*)] \quad (15)$$

As a weighted sum of independent standard exponential variables, V has a general Erlang distribution [14].

$$g(V) = \sum_{i=1}^n \left[\prod_{i \neq k} (\Phi_i - \Phi_k)^{-1} \right] \cdot \Phi_i^{n-2} \cdot \left(\exp \left[\frac{-V}{\Phi_i} \right] \right),$$

$$V \geq 0 \quad (16)$$

The analytic formula for $g(\hat{\theta})$ is now developed, without loss of generality, for the case $n = 2$. Re-expressing (16) using (15),

$$g(V) = \frac{1}{(z_1 \exp[(\theta_0 - \theta^*)] - z_2 \exp[(\theta_0 - \theta^*)])} \cdot$$

$$\left(\exp \left[\frac{-V}{z_1 \exp[z_1(\theta_0 - \theta^*)]} \right] - \exp \left[\frac{-V}{z_2 \exp[z_2(\theta_0 - \theta^*)]} \right] \right),$$

$$V \geq 0 \quad (17)$$

Now, from (14), $g_{[T(\hat{\theta}, \theta^*, w)]}(t) = g_{[V]}(t + \sum z_i)$, and $g_{[T(\hat{\theta}, \theta^*, w)]}(0) = g_{[V]}(\sum z_i)$. Applying equation (2) to (12) and (17), when $V = \sum z_i$,

$$g(\hat{\theta}) = (z_1^2 + z_2^2) \cdot \frac{1}{(z_1 \exp[(\theta_0 - \hat{\theta})] - z_2 \exp[(\theta_0 - \hat{\theta})])} \cdot$$

$$\left(\exp \left[\frac{-(z_1 + z_2)}{z_1 \exp[z_1(\theta_0 - \hat{\theta})]} \right] - \exp \left[\frac{-(z_1 + z_2)}{z_2 \exp[z_2(\theta_0 - \hat{\theta})]} \right] \right), \quad \hat{\theta} \geq 0$$

A program was written to calculate this density and also to construct a simulated probability histogram by deriving samples using a sequence of independent standard exponential random numbers. The MLE $\hat{\theta}$ can be calculated for each simulated data set without difficulty as an analytical formula. For this experiment, $\theta_0 = 0.8$,

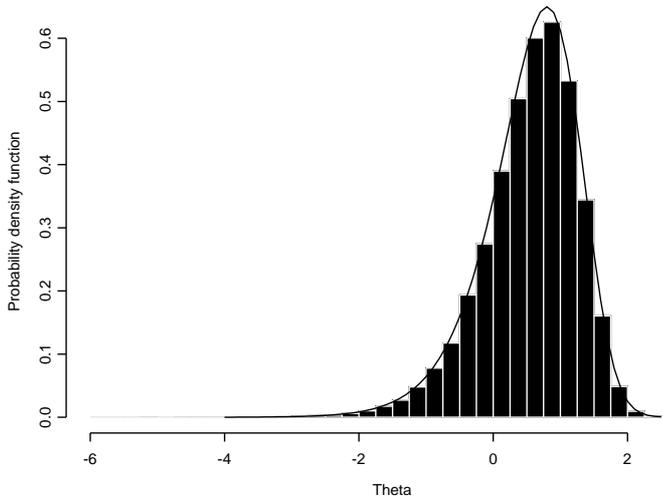


Figure 1:

Exponential regression. Empirical probability density function based on a histogram of estimates from 100,000 simulations, with analytic distribution $g(\hat{\theta})$. (X-axis label θ is $\hat{\theta}$).

$z_1 = 0.75$ and $z_2 = 1.5$. 100,000 simulated sets of data were used. Fig. 1 shows the comparison of the densities derived by the analytic method and by simulation. By inspection, the simulated probability histogram can be seen to agree well with the analytic density. The analytic density agrees well with Fig. 3.1 of [8].

4 Nonlinear regression models

Consider a nonlinear regression model with normal errors, $MN_w(\mu(z); \Sigma)$, where $\mu(z)$ is the $(n \times 1)$ vector of mean responses and Σ is the known $(n \times n)$ covariance matrix. Assume that the data generating model $g_0(w|z)$ has $\mu(z) = f_0(z)$, and the estimation model $g_1(w|\theta, z)$ has $\mu(z) = f(\theta, z)$, where $f(\theta, z)$ can be nonlinear with respect to θ . The estimation model can be restructured in the form (4), with

$$\begin{aligned} a(w) &= w, & b(\theta, z) &= \Sigma^{-1}f(\theta, z), \\ 1^T c(\theta, z) &= -\frac{1}{2}f(\theta, z)^T \Sigma^{-1}f(\theta, z), \\ 1^T d(w) &= -\frac{1}{2}[n \log(2\pi) + \log|\Sigma| + w^T \Sigma^{-1}w] \end{aligned}$$

From (6),

$$T(\hat{\theta}, \theta^*, w) = f'(\theta^*, z) \Sigma^{-1} [w - f(\theta^*, z)]$$

The density of $T(\hat{\theta}, \theta^*, w)$ is as follows.

$$\begin{aligned} MN_T(f'(\theta^*, z) \Sigma^{-1} [f_0(z) - f(\theta^*, z)]; \\ f'(\theta^*, z) \Sigma^{-1} \Sigma \Sigma^{-1} (f'(\theta^*, z))^T) \end{aligned}$$

For the simple iid case $\Sigma = \sigma^2 I$, where I is the $(n \times n)$ identity matrix, the density $g(\hat{\theta})$ can be obtained from (2) as follows.

$$g(\hat{\theta}) = E_w [|j(\theta, w)|_{\theta=\hat{\theta}}] \cdot \left| \frac{2\pi}{\sigma^2} f'(\hat{\theta}, z) (f'(\hat{\theta}, z))^T \right|^{-\frac{1}{2}} \cdot$$

$$\exp \left(-\frac{1}{2\sigma^2} [f(\hat{\theta}, z) - f_0(z)]^T [f'(\hat{\theta}, z)]^T [f'(\hat{\theta}, z) (f'(\hat{\theta}, z))^T]^{-1} [f'(\hat{\theta}, z)] [f(\hat{\theta}, z) - f_0(z)] \right) \quad (18)$$

Further explanation is given by Hingley [11], where there is a recipe to evaluate $E_w [|j(\theta, w)|_{\theta=\hat{\theta}}]$, in particular demonstrating the setup for a two parameter nonlinear model. An example of nonlinear regression is given in Section 5 below.

For the method as developed here, a complete description of the data generating model is assumed known, including specification of σ^2 . However, in an experimental situation, error variance is usually estimated from the residual sum of squares after fitting the model to data, giving an MLE $\hat{\sigma}^2$ that is biased (an unbiased estimate is $\frac{n}{(n-p)} \hat{\sigma}^2$). Now the form of equation (18) shows that $g(\hat{\theta})$ depends on σ^2 but does not depend on $\hat{\sigma}^2$. An extension of the density, to include $\hat{\sigma}^2$ with $\hat{\theta}$ in a joint density $g(\hat{\sigma}^2, \hat{\theta})$, requires multiplication of $g(\hat{\theta})$ by the conditional density $g(\hat{\sigma}^2 | \hat{\theta})$.

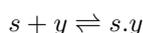
In the common case where the data generating model is a linear model, with a $(n \times p)$ design matrix X_0 and a $(p \times 1)$ parameter vector B_0 , then $f(\theta_0, z) = X_0 B_0$. The density $g(\hat{\sigma}^2 | \hat{\theta})$ is that of a multiple of a central chi-square variate if the estimation model agrees with the data generating model. This distribution does not depend on $f(\hat{\theta}, z)$ or $\hat{\theta}$, so $g(\hat{\sigma}^2 | \hat{\theta}) = g(\hat{\sigma}^2)$ and $g(\hat{\sigma}^2, \hat{\theta}) = g(\hat{\theta}) \cdot g(\hat{\sigma}^2)$. In the case of equivalent nonlinear models, the density $g(\hat{\sigma}^2 | \hat{\theta})$ is also that of a multiple of a variable with a central chi-square distribution. But for either linear or nonlinear models, if there is a distinction between the data generating model and the estimation model, then $g(\hat{\sigma}^2 | \hat{\theta})$ will be related to a noncentral chi-square density, with a noncentrality parameter depending on $f(\hat{\theta}, z) - f_0(z)$ as well as σ^2 and other terms [15]. An effective dependence is created between $g(\hat{\theta})$ and $g(\hat{\sigma}^2)$, in the sense that both depend on $\hat{\theta}$ and $f_0(z)$. However $g(\hat{\sigma}^2, \hat{\theta}) = g(\hat{\theta}) \cdot g(\hat{\sigma}^2 | \hat{\theta})$ can still be calculated if required, and in all cases $g(\hat{\theta})$ can be determined from knowledge of σ^2 without needing to worry about the distribution of $\hat{\sigma}^2$. The above statements apply to the simplest possible formulation of the error process ($\Sigma = \sigma^2 I$) and should be revisited when using models with more intricate error structures.

5 Biochemical titration using a logistic curve (Case 2, estimation model *not equivalent to data generating model*)

This example involves a practical problem that exists in biochemical assays, based on principles of physical chemistry. The assays are titrations and come in various types - such as enzyme-linked immunosorbent assays (ELISA, e.g. [19]), determinations of enzyme kinetics (e.g. [18]) or other binding assays (e.g. [21]). They also relate to methods for total adsorption of substrate onto a heterogenous surface [22], [20]. The measured reaction is an indirect indicator of the way that two or more chemical entities interact. A scientist may postulate several different models for the nature of the interaction, so this is an area where TED can be useful.

In titration experiments, the reactivity of an unknown amount of a component in a chemical preparation is assessed by applying various known concentrations of another substrate substance, with which it reacts. It will be seen that the mathematical form of the function that relates the extent of reaction to the substrate concentration is difficult to fit to data directly. A simpler function can be used for estimation. In the presence of errors in the measured experimental data from the chemical titration experiment, a problem exists of assessing the robustness of estimation of the parameters under the resulting misspecified model. This situation for ELISA estimations of antibody levels in serum is discussed in [19].

Here, the the simplest kind of two component chemical reaction is assumed. The reversible reaction of two substrates can be described by the Law of mass action [26]. Let s and y be the separate chemical components, while $s.y$ is the product of reaction in what is assumed to be a reversible process.



Suppose that a biochemical assay is to be carried out to assess a chemical sample by reacting it with varying concentrations of substrate. From now on, let the terms indicate the concentrations of the reacting components. y_0 and s_{app} are the applied concentrations of the chemical sample and substrate respectively, while $s.y$, s , y are the concentrations respectively of bound chemical sample, unbound chemical sample and substrate that remain at equilibrium.

The affinity of the components for each other can be expressed as an equilibrium constant.

$$K = \frac{s.y}{s.y} = \frac{s.y}{(s_{app} - s.y)(y_0 - s.y)} \quad (19)$$

The magnitude of K represents the propensity of the components to react with each other. If they have a high affinity, then the concentration of product at equilibrium will be high and hence K will have a high value.

The concentration y_0 is fixed but unknown, while the concentration s_{app} is known and allowed to vary. The aim of the exercise is to estimate K and, if possible, y_0 as well. At chemical equilibrium, the fraction of the chemical sample that is bound by substrate is given by a logistic function.

$$f = \frac{s.y}{y + s.y} = \frac{K s y}{y + K s y} = \frac{K s}{1 + K s} \quad (20)$$

It is straightforward to fit this model to data by using an iterative nonlinear estimation routine [10]. But (20) is specified with respect to s , the substrate concentration at equilibrium, rather than the applied concentration of substrate s_{app} . In the usual setup of a titration experiment, a series of readings are taken at different s_{app} values. Equation (20) then does not apply.

Meinert and McHugh [17] give an expression from which the fraction of substrate bound can be found in terms of s_{app} . In the following, the variables from (19) are reparameterised as $\gamma_0 = \log K$, $z = \log[s_{app}]$.

$$f_0(z) = \frac{1}{2y_0} \left[e^z + e^{-\gamma_0} + y_0 - \sqrt{(e^z + e^{-\gamma_0} + y_0)^2 - 4e^z y_0} \right] \quad (21)$$

Assume that a two parameter logistic model is fitted (incorrectly) to applied substrate log concentration z , in the presence of independent homoscedastic normal errors on the assay measurements. This setup can be written as a nonlinear regression model using the formulation that was given in Section 4.

$$w = f(\theta, z) + \epsilon,$$

where ϵ is distributed as $N(0, \sigma^2 I)$, and σ^2 is assumed known.

The function $f_0(z)$ is given by (21) and will now be written as $f_0(\theta_0, z)$, with $\theta_0 = [\gamma_0, y_0]^T$.

The logistic estimation model (20) will be recast as follows.

$$f(\theta_L, z) = \frac{e^{a_L(\gamma_L + z)}}{1 + e^{a_L(\gamma_L + z)}}, \quad (22)$$

where $\theta_L = [\gamma_L, a_L]^T$. Comparison of equation (22) with (20) shows that a_L has been introduced, with $\gamma_L = \frac{1}{a_L} \log K$. The new parameter a_L allows some flexibility in the slope of the fitted function, since the data generating model (21) will be sigmoidal when $f(\theta_0, z)$ is plotted against z , but can not be expected to agree in form with (22). While the aim is to estimate γ_0 , and a_L will be considered as a nuisance parameter in terms of the chemical reaction, in fact there are some contexts in which a_L has a physical meaning. In immunoassays for example, where y represents a heterogenous set of antibody molecules of differing affinity, a_L determines the distribution of affinity [1], [9]. Application of L'Hopital's rule to equation (21), with $2y_0$ being the bottom component, shows

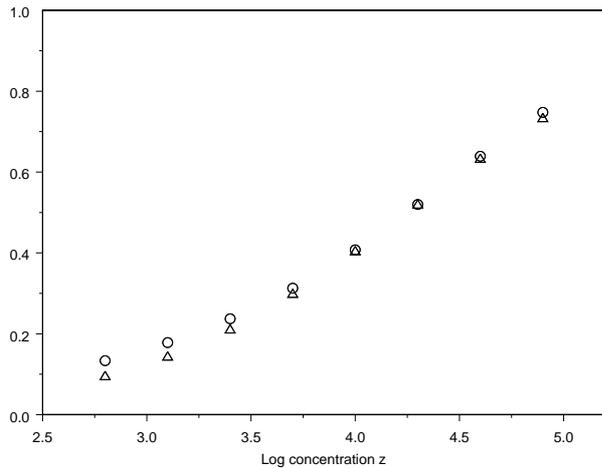


Figure 2:

Fractional saturation curves for biochemical reactions. Triangles: Data generating model $f_0(\theta_0, z)$ according to equation (21); $\gamma_0 = -3$, $y_0 = 100$. Circles: Logistic curve $f(\hat{\theta}_L, z)$ according to equation (22), with $\hat{\theta}_L$ generated using equations (23) and (24), after substituting $\hat{\theta}_0 = \theta_0$.

that, as $y_0 \rightarrow 0$, the curve $f_0(z)$ tends towards the logistic $f(\theta_L, z)$, with $a_L = 1$ and $\gamma_0 = \gamma_L$. Nevertheless, in an assay that contains any chemical sample at all, y_0 will be positive and located away from zero.

This exercise is supported by a set of simulated data using $\gamma_0 = -3$, $y_0 = 100$, and $\sigma^2 = 0.008787$ (comparable to [11], Example 3.4). 16 z values were taken in 8 replicate pairs, that were equally spaced from 2.8 to 4.9. This design is chosen to emulate a typical experimental ELISA setup. Fig. 2 shows $f_0(z)$ using these parameter values. A function $f(\theta_L, z)$ is also shown that is equivalent to $f_0(z)$ according to an *ad-hoc* method that will be described below. It is striking that both plots look similar, even though the physical processes that are assumed to be generating them differ. A significance test to discriminate between the models in the presence of experimental error might have difficulty to do so. The data generating model is slightly asymmetric, with a flattening towards the top of the curve, while the logistic curve is symmetric.

An attempt was made to fit the data generating model (21) separately to each simulated data set. The data sets were made and analysed by an S-Plus program, with the standard iterative nonlinear routine *nls* used to estimate the parameters [25]. The default Gauss-Newton algorithm did not specify derivatives and the starting values each time were γ_0 and y_0 . After successfully fitting five simulated sets of data, the iterative algorithm did not converge for the sixth set and caused a process interrup-

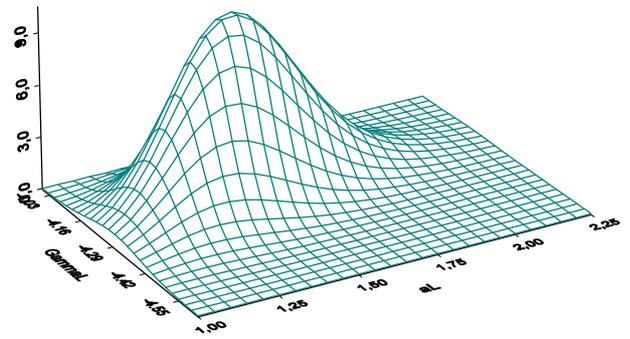


Figure 3:

Distributions of estimates from biochemical titration tests. A probability density function (pdf) surface $g(\hat{\theta})$ according to equation (18). (X-axis label aL is \hat{a}_L . Y-axis label $GammaL$ is $\hat{\gamma}_L$.)

tion. The data set that failed was rather flat compared to the expected values from which it had been generated. It might have been possible to force a fit by using another algorithm or by tailoring the control parameters, but this was not investigated further. It seems that it is difficult to fit the model (21) directly to data.

The TED expression $g(\hat{\theta}_L)$ was calculated from (22), using the method described in [11] for a nonlinear regression model that is based on equation (18), with two estimable parameters and homoscedastic independent normal errors. The fit of (22) was assessed on a series of simulated data sets that were generated by (21) in the same way as described above, using as starting values for the fitting algorithm $\gamma_L = -3$, $a_L = 1.5$. This time 1,000 sets of simulated data were fitted without apparent difficulty, although experiments with much larger numbers of simulations again suggested that there can be occasional occurrences of non convergence.

Figs 3 and 4 show a comparison of $g(\hat{\theta}_L)$ with an empirical density plot of the 1,000 results from the simulated data sets. Visual inspection indicates agreement of the simulated data with the analytic density. Bias can exist in any situation where the data generating model or the estimation model are asymmetric and the experimental design does not centre around the mid point of the model. Here, both the analytic density and the simulations demonstrate that $\hat{\gamma}_L$ is indeed a biased estimator of γ_0 . The distribution is centred at about $\hat{\gamma}_L = -4.22$ (compared to $\gamma_0 = -3$), $\hat{a}_L = 1.45$, with positive correla-

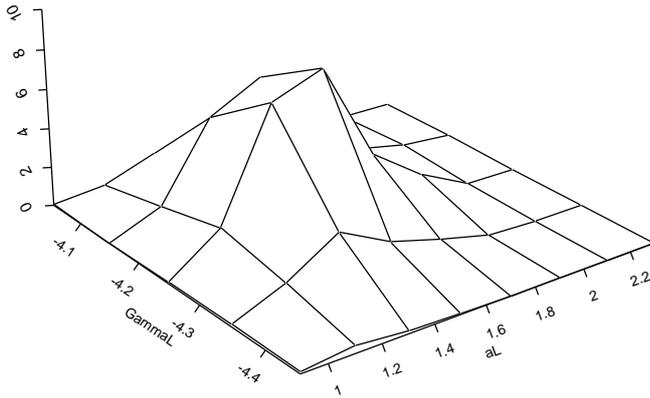


Figure 4:

Distributions of estimates from biochemical titration tests. Empirical probability density function (pdf) based on a histogram of estimates from 1,000 simulation experiments. (X-axis label aL is \hat{a}_L . Y-axis label $\text{Gamma}L$ is $\hat{\gamma}_L$.)

tion between the estimates.

How should the experimenter proceed to estimate γ_0 after fitting a logistic curve to a single set of data? No perfect solution will be offered here, but a suggestion can be made for a correction that reduces the bias of the estimate to some extent. A reparameterisation of the estimates can be made from $\hat{\theta}_L = [\hat{\gamma}_L, \hat{a}_L]^T$ to $\hat{\theta}_0 = [\hat{\gamma}_0, \hat{y}_0]^T$, by using an *ad hoc* method to pick an equivalent curve $f_0(\hat{\theta}_0, z)$ to the fitted curve $f(\hat{\theta}_L, z)$. Let $z_{0.5}$ be the value of z at half saturation. Set $\hat{\gamma}_0$ to give $f_0(\hat{\theta}_0, z)$ with the same half saturation value $z_{0.5}$ and slope $\frac{\partial f}{\partial z} |_{z=z_{0.5}}$. This is chosen because the algebra is simple and because a well designed experiment will centre measurements roughly around $z_{0.5}$. Equation (22) gives $\gamma_L = -z_{0.5}$, and $\frac{\delta f(\hat{\theta}_L, z)}{\delta z} |_{z=z_{0.5}} = \frac{a_L}{4}$, leading to the following suggestions for corrections.

$$\hat{\gamma}_0 = \hat{\gamma}_L + \log\left(\frac{4 - \hat{a}_L}{\hat{a}_L}\right) \quad (23)$$

$$\hat{y}_0 = 4e^{-\hat{\gamma}_L} \left(1 - \frac{\hat{a}_L}{2}\right) \quad (24)$$

Constraints on \hat{a}_L are suggested by empirical equation (23) as $0 < \hat{a}_L < 4$, and by empirical equation (24) as $\hat{a}_L < 2$, implying $0 < \hat{a}_L < 2$. However, in the simulations a few \hat{a}_L values are above 2, which demonstrates that this is indeed an approximate argument (\hat{a}_L : mean = 1.45, min. = 0.85, max. = 2.25; $\hat{\gamma}_L$: mean = -4.22, min. = -4.49, max. = -4.01).

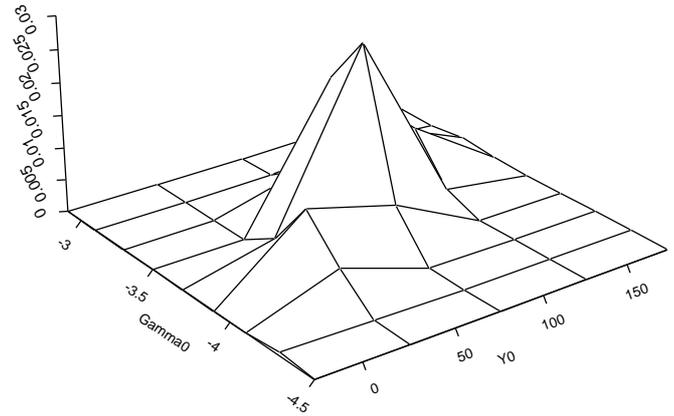


Figure 5:

Distributions of estimates from biochemical titration tests. Empirical probability density function (pdf) based on a histogram of estimates from 1,000 simulation experiments, after reparameterisation using the bias reducing transformation according to equations (23) and (24). (X-axis label Y_0 is \hat{y}_0 . Y-axis label $\text{Gamma}0$ is $\hat{\gamma}_0$.)

Fig. 5 shows an empirical density of the same 1,000 simulation results that were described in Fig. 4, after transformation of the parameter estimates from each set of simulated data from $\hat{\theta}_L$ to $\hat{\theta}_0$. The reparameterisation has reduced the bias in estimation of γ_0 (-3), and now also gives information about the useful parameter y_0 (100) (\hat{y}_0 : mean = 75.5, min. = -29.3, max. = 184; $\hat{\gamma}_0$: mean = -3.65, min. = -4.40, max. = -2.92). The distribution of \hat{y}_0 is diffuse - although it covers y_0 (100), it also reaches below 0 which has no physical meaning in terms of applied concentrations for the chemical reaction. Recall that equation (21) showed that γ_0 is a measure of the reactivity of the substance under assay, while y_0 indicates the total amount of the substance. The plateau of the titration curve in some assays is not unity but is proportional to the amount of substance under test [13]. The plateau thus gives additional information on y_0 that could be combined with the estimate \hat{y}_0 from the two parameter logistic estimation. The model will not be extended in this way here.

The density of the corrected estimates, and other potential corrections, could also be investigated using TED by obtaining the density of the transformed variables after multiplication by a Jacobian as follows.

$$g(\hat{\theta}_0) = |d\hat{\theta}_L/d\hat{\theta}_0| \cdot g(\hat{\theta}_L)$$

6 Uses and extensions of TED

TED is a framework approach for determining the density of the MLE, where the functional forms of both the data generating model and the estimation model are distinctly specified. When models are the same, alternative analytic approximate and exact methods are available [12]. But apparently no other analytic techniques exist when the models differ. Tractability can be achieved as long as the density $g_{[T(\hat{\theta}, \theta^*, w)]}(t)$ can be determined and, for the case of curved exponential family models, this involves the determination of the density of the linear transform (6) of a functional of the data. This can be done for simple estimation models but there may be cases where it is hard or impossible. Then approximations to $g_{[T(\hat{\theta}, \theta^*, w)]}(t)$ can be tried and incorporated into the framework.

In the frequentist mode, if common parameters exist in the data generating model and the estimation model, then TED can be used to construct a formal Robustness Index (RI) for the comparison of the adequacy of an estimation model with a data generating model on a particular experimental design [11]. The emphasis is then on the estimation of equivalent parameters on a set of candidate models that have a physical interpretation in terms of an underlying experiment. This will often be more relevant to the scientist than the rather sterile practice of testing a null hypothesis vs. an alternative hypothesis to discriminate between models. An advantage of TED is to provide an analytic handle in an area where otherwise simulations are used. The RI approach with TED may itself require numerical integration to be carried out. It is not suggested that numerical approaches are inappropriate, only that TED enables the departure point for simulation work to be put further down the line where more interesting results may be obtained.

In the Bayesian framework, an experimental approach might be to align the prior density of a parameter θ on a particular model with the TED density $g(\hat{\theta})$, since this incorporates the range of variability in θ that might exist when data are incorporated in order to calculate a posterior density. Applications in Bayesian model selection and in Bayesian model averaging might also be possible [2], [3].

Model uncertainty can be countered by using distribution free techniques and by robust estimation methods. Data contamination, systematic error and various other forms of heteroscedasticity are all facts of life that can also be considered by using TED.

Theoretical extensions could be considered to the basic design concept of TED. Other choices are possible for $T(\theta, \theta^*, w)$ for equation (1) in order to develop a method to find $g(\hat{\theta})$. Such expressions should also involve a constant like θ^* and an unspecified parameter θ . But they would probably have to be more complicated than the

existing definition in (1). It seems to be inadequate to use the simpler expression $T(\theta^*, w) = l'(\theta^*, w)$, since the absence of θ means that the observed information can not be found by differentiation. It is also inappropriate to use $T(\hat{\theta}, w) = l'(\theta, w)$ by itself, since then the equivalent to $T(\hat{\theta}, \theta^*, w)$ is $l'(\theta, w)|_{\theta=\hat{\theta}}$, which is necessarily 0 and so $g_{[l'(\hat{\theta}, w)]}(l')$ is degenerate.

There seems to be a need to restrict θ to be some calculated statistic of the data. Otherwise it is necessary to develop a density for $T(\theta, \theta^*, w)$ itself, a mixture of the terms involving w and θ . Transformation of the known density $g_0(w)$ to that for an unspecified θ is probably not feasible. For example, it would be quite difficult to find the density for the equivalent to expression (6) for curved exponential families with θ unspecified.

$$T(\theta, \theta^*, w) =$$

$$[b'(\theta^*, z) \cdot (a(w) - f(\theta^*, z))] - [b'(\theta, z) \cdot (a(w) - f(\theta, z))]$$

TED could be expanded to cover other estimators than the MLE, although the algebra may be less convenient. One possible extension is to minimum contrast estimators, in order to complement the approach of Skovgaard [23] when models differ. Similar derivations to that of equation (2) can be applied. Another application is to find the density $g(\hat{\theta})$ for a MLE that assumes no contamination when in fact some degree of prespecified contamination does occur in the data generating model. One way to assess robustness properties is via the influence function [6], to which the analytic function $g(\hat{\theta})$ should be able to give some support.

The TED approach could also be generalised to robust regression using M-estimators. A simple example is to find the mean of a normal sample. Consider an example in [12] which assumes that σ^2 is known and simplify further to $\sigma^2 = 1$. An extended version of TED can be imagined, where T is based on a robustified equivalent to the score function $l'(\theta, w)$.

A Huber function $\Psi(\theta, w)$ can to be used to obtain a robust estimate $\hat{\mu}_m$ of the mean μ [24].

$$\Psi(\mu_m, w_i) =$$

$$\frac{w_i - \mu_m}{0.6745},$$

$$\text{if } \left| \frac{w_i - \mu_m}{0.6745} \right| \leq 1.28$$

$$1.28 \text{sign}(w_i - \mu_m),$$

$$\text{if } \left| \frac{w_i - \mu_m}{0.6745} \right| > 1.28$$

In this case, define $T(\mu_m, \mu_m^*, w) = \sum_{i=1}^n t_i(\mu_m, \mu_m^*, w_i)$, where

$$t_i(\mu_m, \mu_m^*, w_i) = \frac{w_i - \mu_m^*}{0.6745} - \frac{w_i - \mu_m}{0.6745},$$

if $\left| \frac{w_i - \mu_m^*}{0.6745} \right| \leq 1.28, \left| \frac{w_i - \mu_m}{0.6745} \right| \leq 1.28$

$$\frac{w_i - \mu_m^*}{0.6745} - 1.28 \operatorname{sign}(w_i - \mu_m),$$

if $\left| \frac{w_i - \mu_m^*}{0.6745} \right| \leq 1.28, \left| \frac{w_i - \mu_m}{0.6745} \right| > 1.28$

$$1.28 \operatorname{sign}(w_i - \mu_m^*) - \frac{w_i - \mu_m}{0.6745},$$

if $\left| \frac{w_i - \mu_m^*}{0.6745} \right| > 1.28, \left| \frac{w_i - \mu_m}{0.6745} \right| \leq 1.28$

$$1.28 \operatorname{sign}(w_i - \mu_m^*) - 1.28 \operatorname{sign}(w_i - \mu_m),$$

if $\left| \frac{w_i - \mu_m^*}{0.6745} \right| > 1.28, \left| \frac{w_i - \mu_m}{0.6745} \right| > 1.28$

The problem is to determine the distribution of T , either on a data generating model that is normal or, preferably, on a suitably contaminated normal for which robust estimation is worthwhile.

It remains to be investigated whether these kinds of robust estimates are appropriate when studying problems of model misspecification. Perhaps robustness can be examined better by looking directly at the influence functions for various types of estimates and model combinations, with $g(\hat{\theta})$ merely giving a confirmation in each case.

7 Conclusion

The exposition in this paper has been given with a view to assist with practical modelling problems. The examples in Sections 3 and 5 can be extended and, when real data are available, an information criterion for comparing the adequacy of competing models can be used [12]. Regression techniques are applied in most practical statistical modelling studies and the possibilities for further applications of TED therefore seem to be almost unlimited.

References

[1] Bell, G., "Mathematical model of clonal selection and antibody affinity", *Journal of theoretical biology*, V29, pp. 191-232, 1970.

[2] Claeskens, G. and Hjort, N.L., *Model selection and model averaging*, Cambridge, 2008.

[3] Davison, A.C., *Statistical models*, Cambridge, 2003.

[4] Dobson, A.J., *An introduction to statistical modelling*, London, Chapman and Hall, 1983.

[5] Favero, C.A., *Applied Macroeconometrics*, Oxford, 2001.

[6] Hampel, F.R., Ronchetti, E.W., Rousseeuw, P.J. and Stahel, W.A., *Robust statistics, the approach based on influence functions*, New York, Wiley, 1986.

[7] Hillier, G. and Armstrong, H., "The density of the maximum likelihood estimator", *Econometrica*, V67, 1459-1470, 1999.

[8] Hillier, G. and O'Brien, R., "Exact properties of the maximum likelihood estimator in exponential regression models: a differential geometric approach", in *Applications of differential geometry to econometrics*, Cambridge, pp. 85-118, 2000.

[9] Hingley, P.J., "A clonal selection based timecourse model for antibody responses to killed vaccine, with applications to foot and mouth disease", *Biometrics*, V47, pp. 1019-1047, 1991.

[10] Hingley, P.J., "Nonlinear regression modelling in APL", *Vector*, V9, pp. 109-122, 1992.

[11] Hingley, P.J., "Analytic estimator densities for common parameters under misspecified models", *Statistics for Industry and Technology*, pp. 119-130, 2004.

[12] Hingley, P.J., "Distributions of maximum likelihood estimators and model comparisons", in *Current themes in Engineering Science*, American Institute of Physics, pp. 111-122, 2008.

[13] Hingley, P.J. and Ouldrige, E.J., "The use of a logistic model for the quantitative interpretation of indirect sandwich enzyme labelled immunosorbent assays (ELISA) for antibodies and antigens in foot and mouth disease", *Computers in Biology and Medicine*, V15, pp. 137-152, 1985.

[14] Johnson, N.L., Kotz, S. and Balakrishnan, N., *Continuous univariate distributions, Volume 1*, New York, Wiley, 1994.

[15] Johnson, N.L., Kotz, S. and Balakrishnan, N., *Continuous univariate distributions, Volume 2*, New York, Wiley, 1995.

[16] Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E. and Getz, W.M., "Superspreading and the effect of individual variation on disease emergence", *Nature*, V438 pp. 355-359, 2005.

[17] Meinert, C.L. and McHugh R.B., "The biometry of an isotope displacement immunologic microassay", *Mathematical Biosciences*, V2 pp. 319-338, 1968.

[18] Murphy, E.F., Gilmour, S.G. and Crabbe, M.J.C., "Efficient and accurate experimental design for enzyme kinetics: Bayesian studies reveal a systematic approach", *Journal of Biochemical and Biophysical Methods*, V55, pp. 155-178, 2003.

- [19] Ouldridge, E.J., Barnett, P.V., Hingley, P.J. and Rweyemamu, M.M., "The differentiation of foot and mouth disease virus strains using an indirect sandwich enzyme-linked immunosorbent assay saturation model", *Journal of Biological Standardization*, V12 pp. 367-377, 1984.
- [20] Ross, S. and Oliver, S.P., *On physical adsorption*, New York, Wiley, 1964.
- [21] Russ, W.P., Lowery, D.M., Mishra, P., Yaffe, M.B. and Ranganathan R., "Natural-like function in artificial WW domains", *Nature*, V437 pp. 579-583, 2005.
- [22] Sips, R., "On the structure of a catalyst surface", *Journal of Chemical Physics*, V16 pp. 490-495, 1948.
- [23] Skovgaard, I.M., "On the density of minimum contrast estimators", *Annals of Statistics*, V18, pp. 779-789, 1990.
- [24] Wilcox, R.R., *Introduction to robust estimation and hypothesis testing, Second edition*, Burlington, Elsevier, pp. 73-79, 2005.
- [25] Venables, W.N. and Ripley, B.D., *Modern applied statistics with S*, New York, Springer, 2002.
- [26] Yudkin, M and Offord, R., *Comprehensible biochemistry*, London, Longman, 1973.