Corrections for the PhD thesis
Bias Reduction in Exponential Family Nonlinear Models

Ioannis Kosmidis

14th January 2009

Corrections

1. Theorem 3.7.1 in subsection 3.7.4 is wrong. The contents of Subsection 3.7.4 should be discarded and replaced by the following:

   "In the case of general exponential families, we have derived a general necessary and sufficient condition for the existence of a penalized likelihood corresponding to the modified scores based on the expected information (see Subsection 3.2.3). The derivation was based on the re-expression of the modified scores as ordinary scores plus derivative of the logarithm of Jeffreys prior plus an extra term. This extra term is zero for canonical families and so the modified scores correspond to penalization of the likelihood by Jeffreys prior.

   Here we present a more specialized but not less important result within the class of univariate GLMs. The following theorem identifies those non-canonical link functions which always — i.e, regardless of the dimension or structure of X — do admit such a penalized likelihood interpretation.

   Theorem 3.7.1: *In the class of univariate generalized linear models, there exists a penalized log-likelihood l^* such that $\nabla l^*(\beta) \equiv U(\beta) + A^{(E)}(\beta)$, for all possible specifications of design matrix X, if and only if the inverse link derivatives $d_r = 1/g_r'(\mu_r)$ satisfy*

   $$d_r \equiv \alpha_r \kappa_{2,r}^{\omega} \quad (r = 1, \ldots, n)\,, \tag{3.28}$$

   where α_r $(r = 1, \ldots, n)$ and ω do not depend on the model parameters. When condition (3.28) holds, the penalized log-likelihood is

   $$l^*(\beta) = \begin{cases} l(\beta) \quad + \quad \dfrac{1}{4}\displaystyle\sum_r \log \kappa_{2,r}(\beta)^{h_r} & (\omega = 1/2) \\[2ex] l(\beta) \quad + \quad \dfrac{\omega}{4\omega - 2} \log |F(\beta)| & (\omega \neq 1/2)\,. \end{cases} \tag{3.29}$$

   Proof. Note that $d_r'/d_r = \mathrm{d}\log d_r / \mathrm{d}\eta_r$ and so, for adjustments based on the expected information, (3.22) can be written as

   $$U_t^* = U_t + \frac{1}{2}\,\mathrm{trace}\,(HET_t) \quad (t = 1, \ldots, p)\ , \tag{3.30}$$

   with $E = \mathrm{diag}(\mathrm{d}\log d_1/\mathrm{d}\eta_1, \ldots, \mathrm{d}\log d_n/\mathrm{d}\eta_n)$ and $T_t = \mathrm{diag}(x_{1t}, \ldots, x_{nt})$. As, for example, in the case of the existence of quasi-likelihoods in McCullagh & Nelder (1989, § 9.3.2), there exists a penalized log-likelihood l^* such that $\nabla l^*(\beta) \equiv U(\beta) + A^{(E)}(\beta)$, if and only if $\partial U_s^*(\beta)/\partial\beta_t = \partial U_t^*(\beta)/\partial\beta_s$ for every $s, t \in \{1, 2, \ldots, p\}$, $s \neq t$. By (3.30), this holds if and only if

   $$\frac{\partial \mathrm{trace}\,(HET_t)}{\partial\beta_s} = \mathrm{trace}\left(H \frac{\partial E}{\partial\beta_s} T_t\right) + \mathrm{trace}\left(\frac{\partial H}{\partial\beta_s} E T_t\right)$$

   is invariant under interchange of the subscripts s and t. The first term in the right hand side of the above expression is invariant because $\partial E/\partial\beta_s = ET_s$ and T_s and T_t are by definition diagonal so that matrix multiplication is commutative for them. For the second term,

$$\frac{\partial H}{\partial \beta_s} = -X(X^TWX)^{-1}X^TW_sX(X^TWX)^{-1}X^TW + X(X^TWX)^{-1}X^TW_s\,,$$

where $W_s = \partial W/\partial \beta_s = W(2E - \Lambda)T_s$ with
$\Lambda = \mathrm{diag}(\mathrm{dlog}\,\kappa_{2,1}/\mathrm{d}\eta_1, \ldots, \mathrm{dlog}\,\kappa_{2,n}/\mathrm{d}\eta_n)$. Thus,

$$\begin{aligned}\mathrm{trace}\left(\frac{\partial H}{\partial \beta_s}ET_t\right) =& 2\,\mathrm{trace}\,(HET_sET_t) - 2\,\mathrm{trace}\,(HET_sHET_t) - \\ & \mathrm{trace}\,(H\Lambda T_sET_t) + \mathrm{trace}\,(H\Lambda T_sHET_t)\ .\end{aligned}$$

By the properties of the trace function, the first three terms in the right hand side of the above expression are invariant under interchange of s and t. Thus the condition is reduced to the invariance of $\mathrm{trace}(H\Lambda T_sHET_t)$. The projection matrix H can be written as $H = SW$, with $S = X(X^TWX)^{-1}X^T$. Let $\tilde{H} = W^{1/2}SW^{1/2}$. Taking into account the symmetry of $\tilde{H}$, some trivial algebra and an application of Theorem 3 of Magnus & Neudecker (1999, Chapter 2) gives

$$\mathrm{trace}\,(H\Lambda T_sHET_t) = \mathrm{trace}\left(\tilde{H}T_s\Lambda\tilde{H}ET_t\right) = (\mathrm{vec}\,T_t)^T\left\{\left(E\tilde{H}\Lambda\right)\otimes\tilde{H}\right\}\mathrm{vec}\,T_s\,. \qquad (3.31)$$

The columns of X are by assumption linearly independent and thus (3.31) is invariant under interchanges of s and t if and only if $a^T\{(E\tilde{H}\Lambda)\otimes\tilde{H}\}b$ is a symmetric bilinear form for distinct vectors a and b of appropriate dimension, or equivalently if and only if $E\tilde{H}\Lambda$ is symmetric, i.e.,

$$\frac{\mathrm{dlog}\,d_r}{\mathrm{d}\eta_r}\frac{\mathrm{dlog}\,\kappa_{2,i}}{\mathrm{d}\eta_i}\tilde{h}_{ri} = \frac{\mathrm{dlog}\,\kappa_{2,r}}{\mathrm{d}\eta_r}\frac{\mathrm{dlog}\,d_i}{\mathrm{d}\eta_i}\tilde{h}_{ri} \quad (r, i = 1, \ldots, n; r > i)\,, \qquad (3.32)$$

with $\tilde{h}_{ri}$ the (r, i)th element of $\tilde{H}$.

In general the above equations are not satisfied simultaneously, except possibly for special structures of the design matrix X, which cause $\tilde{h}_{ri} = 0$ for a set of pairs (r, i). Hence, assuming that $\tilde{h}_{ri} \neq 0$ $(r, i = 1, \ldots, n; r > i)$, the general equation in (3.32) reduces to $\mathrm{dlog}\,d_r/\mathrm{d}\eta_r\mathrm{dlog}\,\kappa_{2,i}/\mathrm{d}\eta_i = \mathrm{dlog}\,\kappa_{2,r}/\mathrm{d}\eta_r\mathrm{dlog}\,d_i/\mathrm{d}\eta_i$. Now, if $\mathrm{dlog}\,\kappa_{2,r}/\mathrm{d}\eta_r = \mathrm{dlog}\,\kappa_{2,i}/\mathrm{d}\eta_i = 0$ for some pair (r, i) then the equation for this (r, i) is satisfied. Thus, without loss of generality assume that $\mathrm{dlog}\,\kappa_{2,r}/\mathrm{d}\eta_r \neq 0$ for every $r \in \{1, \ldots, n\}$. Under these assumptions condition (3.32) can be written as $\mathrm{dlog}\,d_r/\mathrm{d}\eta_r = \omega\mathrm{dlog}\,\kappa_{2,r}/\mathrm{d}\eta_r$ $(r = 1, \ldots, n)$, where ω does not depend on the model parameters. By integration of both sides of $\mathrm{dlog}\,d_r/\mathrm{d}\eta_r = \omega\mathrm{dlog}\,\kappa_{2,r}/\mathrm{d}\eta_r$ with respect to η_r, a necessary condition for the adjusted score to be the gradient of a penalized likelihood is thus

$$d_r \equiv \alpha_r\kappa_{2,r}^{\omega} \quad (r = 1, \ldots, n)\ , \qquad (3.33)$$

where $\{\alpha_r : r = 1, \ldots, n\}$ are real constants not depending on the model parameters.

To check that (3.33) is sufficient, simply note that if we substitute accordingly, the matrix $E\tilde{H}\Lambda$ is symmetric.

In addition, if condition (3.33) is satisfied for some ω and α_r ($r = 1, \ldots, n$) then the rth diagonal element of E is $\mathrm{d}\log d_r/\mathrm{d}\eta_r = \omega\kappa'_{2,r}/\kappa_{2,r}$ for every $r \in \{1, \ldots, n\}$, with $\kappa'_{2,r} = \mathrm{d}\kappa_{2,r}/\mathrm{d}\eta_r$. On the other hand, $\mathrm{d}w_r/\mathrm{d}\eta_r = (2\omega - 1)w_r\kappa'_{2,r}/\kappa_{2,r}$. Hence, by (3.30) and for $\omega \neq 1/2$ the tth component of the adjusted score vector is

$$U_t(\beta) + \frac{\omega}{4\omega - 2}\,\mathrm{trace}\left\{X\left(X^T W(\beta) X\right)^{-1} X^T W_t(\beta)\right\} = \frac{\partial}{\partial\beta_t}\left\{l(\beta) + \frac{\omega}{4\omega - 2}\log\left|X^T W(\beta) X\right|\right\},$$

where and $W_t(\beta) = \partial W(\beta)/\partial\beta_t$.

If $\omega = 1/2$ then $w_r = \alpha_r^2$ ($r = 1, \ldots, n$). Hence, the hat matrix H does not depend on the model parameters. Thus, by (3.22), the tth component of the adjusted score vector is

$$U_t(\beta) + \frac{1}{4}\sum_r h_r \frac{\kappa'_{2,r}(\beta)}{\kappa_{2,r}(\beta)} x_{rt} = \frac{\partial}{\partial\beta_t}\left\{l(\beta) + \frac{1}{4}\sum_r \log\kappa_{2,r}(\beta)^{h_r}\right\}.$$

□

In the above theorem the canonical link is the special case $\omega = 1$. With $\omega = 0$, condition (3.28) refers to identity links for which the log-likelihood penalty is identically zero. The case $\omega = 1/2$ is special on account of the fact that the working weights, and hence F and H, do not in that case depend on β.

Example 3.7.1: Consider a Poisson generalized linear model with link function from the power family $\eta = (\mu^\nu - 1)/\nu$ (McCullagh & Nelder, 1989, § 2.2.3). Then $d_r = \mu_r^{1-\nu}$ and $\kappa_{2,r} = \mu_r$ ($r = 1, \ldots, n$). Bias reduction based on expected information is equivalent to maximization of penalized likelihood (3.29) with $\omega = 1 - \nu$.

Theorem 3.7.1 has direct practical consequences, notably for the construction of confidence sets. The use of profile penalized likelihood as suggested for example in Heinze & Schemper (2002) and Bull et al. (2007) is always available for a generalized linear model whose link function satisfies condition (3.28), but typically is not possible otherwise. Models which fail to meet condition (3.28) include, for example, probit and complementary log-log models for binary responses."

2. On page 69, in the last paragraph of Subsection 5.2.1:

   "We should mention that in contrast to the case of logistic regression, Theorem 3.7.1 shows that there does not exist a penalized likelihood corresponding to either (5.2), (5.3), or (5.4)."

   should be replaced by

"We should mention that in contrast to the case of logistic regression, Theorem 3.7.1 can be used to show that generally there does not exist a penalized likelihood corresponding to either (5.2), (5.3), or (5.4), except for very special structures of matrix X like, for example, the one-way layout."

3. On page 84, in the last paragraph of Subsection 5.2.6:

   "However, as Theorem 3.7.1 shows, there is no penalized likelihood corresponding to non-canonical models within the class of GLMs. Thus, for the BR estimates, there is no argument motivating the use of confidence intervals based on penalization of the likelihood by Jeffreys prior."

   should be replaced by

   "By Theorem 3.7.1, in the case of binomial response GLMs, a penalized likelihood corresponding to the modified scores does not generally exist for non-canonical links."

4. on page 103, second paragraph:

   "We have focused on univariate GLMs, where despite the fact that penalized likelihoods exist only for canonical links, we have shown ..."

   should be replaced by

   "We have focused on univariate GLMs, where we have identified the models that admit a penalized likelihood interpretation of the bias-reduction method with modifications based on the expected information and we have shown ..."

5. On page 26, expression (3.16), $W_{rs} \otimes 1_q$ should be $W_{rs} \otimes 1_p$. The same typo should be corrected in the displayed expression for $P_t + Q_t$ before expression (3.16). Corresponding corrections should also be applied to the relevant parts in (B.7), the expression after (B.7), (B.8), (B.9) and in the algebraic derivation on page 125. That is $[D_r^T]_s \otimes 1_q$ should be $[D_r^T]_s \otimes 1_p$, $\mathcal{D}(\mu_{rs}; \eta_r) \otimes 1_q$ should be $\mathcal{D}(\mu_{rs}; \eta_r) \otimes 1_p$ and $W_{rs} \otimes 1_q$ should be $W_{rs} \otimes 1_p$.

Acknowledgements

The author is grateful to Professor David Firth for bringing to the author's attention that Theorem 3.7.1 is wrong and for his advice on the derivation of the correct result.

Bibliography

Bull, S. B., Lewinger, J. B. & Lee, S. S. F. (2007). Confidence intervals for multinomial logistic regression in sparse data. *Statistics in Medicine* **26**, 903–918.

Heinze, G. & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409–2419.

Magnus, J. R. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester: Wiley.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. London: Chapman and Hall, 2nd ed.