# Exact gradient and Hessian for TOPALS model fitting with age-grouped input data

Carl Schmertmann

October 2021

## 1 Main Idea

Here I show the derivation for the analytical gradient and Hessian of the penalized likelihood with respect to TOPALS model parameters. These expressions allow optimization by Newton-Raphson iteration, which is slightly different from the iteratively reweighted least squares approach. In particular, the covariance matrix estimated from the IRLS approach at the final estimate isn't *quite* the same as the approximation that uses the inverse of the negative of the Hessian.

### 1.1 Objective

We wish to fit a TOPALS model mortality schedule for $A$ single years of age $x = 0, 1, \ldots, (A-1)$ from exposure and death data for $G \leq A$ age groups: $N_1 \ldots N_G$ and deaths $D_1 \ldots D_G$. We include a very small roughness penalty, which will affect the fit only when there are age groups with zero exposure.

### 1.2 Sample Data

Observed data consists of deaths $D_g$ and exposure $N_g$ for closed, non-overlapping age groups $g = 1 \ldots G$. Age groups are defined by a vector of $G+1$ boundary ages – e.g. if the boundary ages are $0, 1, 5, 10, \ldots, 85, 90$ then the age groups are $[0, 1), [1, 5), \ldots, [85, 90)$. Denote $X_g$ as the set of integer ages that belong to group $g$, and $n_g$ as the number of ages in $X_g$.

Typical age groups have boundaries like $0, 1, 5, 10, \ldots, 85, 90$. However, if single-year death and exposure data is available then the "groups" could be simply integer ages, in which case boundaries for $A$ single-year "groups" would be $0, 1, 2, \ldots, A$

### 1.3 Poisson Log Likelihood for Grouped Data

Regardless of the age grouping in the data, we assume that there is a latent schedule for $A$ single-year ages. Deaths at each single-year age have a Poisson

distribution with expected value equal to the (possibly unobserved) exposure $N_x$ times the mortality rate:

$$D_x \sim Poisson\,(N_x\,\mu_x) \quad x = 0, 1, 2, \ldots (A-1)$$

Under the standard assumption that age-specific deaths are statistically independent, this implies that deaths are also Poisson distributed within each age group:

$$D_g \sim Poisson\left(\sum_{x \in X_g} N_x\,\mu_x\right) \quad g = 1, 2, \ldots G$$

Expressing this in terms of the total *observed* exposure in each age group,

$$D_g \sim Poisson\left(N_g \sum_{x \in X_g} \left[\frac{N_x}{N_g}\right] \mu_x\right)$$

or

$$D_g \sim Poisson\,(N_g\,M_g)$$

where $M_g$ represents the exposure-weighted average mortality rate across ages in group $g$.

In the absence of single-year exposure data, we approximate with $M_g = [\frac{1}{n_g}] \sum_{x \in X_g} \mu_x$.[1]
The entire vector of $G$ averaged mortality rates by age group is therefore

$$\boldsymbol{M} = \begin{pmatrix} n_1^{-1} \ldots n_1^{-1} & \cdots & 0 \ldots 0 \\ \vdots & \ddots & \vdots \\ 0 \ldots 0 & \cdots & n_G^{-1} \ldots n_G^{-1} \end{pmatrix} \begin{pmatrix} \mu_0 \\ \vdots \\ \mu_{A-1} \end{pmatrix} = \boldsymbol{W\mu}$$

where $\boldsymbol{W}$ is a $G \times A$ matrix of weights with row sums all equal to one.[2]

The sample log likelihood for a single-year mortality schedule $\boldsymbol{\mu}$ is thus

$$\begin{aligned} \ln L(\boldsymbol{\mu}) &= c + \sum_{g=1}^{G} (\, D_g \ln M_g - N_g M_g \,) \\ &= c + \sum_{g=1}^{G} \left( D_g \ln M_g - \hat{D}_g \right) \end{aligned}$$

where $\hat{D}_g = N_g M_g$ represents the expected number of deaths in group $g$. It will also be useful to express the likelihood in matrix terms:

$$\ln L = c + [\ln M_1 \cdots \ln M_G]\,\boldsymbol{D} - \hat{\boldsymbol{D}}'\mathbf{1} \tag{1}$$

where $\boldsymbol{D}$ and $\hat{\boldsymbol{D}}$ are $G \times 1$ vectors of observed and expected deaths, ordered by age group, and $\mathbf{1}$ is a $G \times 1$ vector of ones.

[1] An alternative approach might interpolate single-year exposure $(N_0, N_1, \ldots N_{A-1})$ from the available $N_g$ to construct more nuanced weights for intra-group averaging of mortality rates. In most populations this would have only minor effects on estimates.

[2] When "groups" correspond to single years, $\boldsymbol{W} = \boldsymbol{I}_A$ and $\boldsymbol{M} = \boldsymbol{\mu}$.

## 1.4 TOPALS Model Mortality Schedule

The TOPALS model for single-year mortality rates is

$$\boldsymbol{\lambda} = \boldsymbol{\lambda}^* + \boldsymbol{B}\boldsymbol{\alpha}$$

where $\boldsymbol{\lambda}$ is an $A \times 1$ vector of age-specific log mortality rates for ages $0 \dots (A-1)$, $\boldsymbol{\lambda}^*$ s a vector with fixed constants representing a standard schedule for those ages, $\boldsymbol{B}$ is a $A \times K$ matrix of linear B-spline constants[3] , and $\boldsymbol{\alpha}$ is a $K$-dimensional vector representing deviations from the standard log mortality schedule at specified ages.

In this model the log mortality rate at age $x \in \{0 \dots (A-1)\}$ is

$$\ln \mu_x = \lambda_x = \lambda_x^* + \boldsymbol{b}_x' \boldsymbol{\alpha}$$

and the mortality rate is

$$\mu_x = exp\left(\lambda_x^* + \boldsymbol{b}_x' \boldsymbol{\alpha}\right)$$

where $\boldsymbol{b}_x'$ is the $1 \times K$ row of $\boldsymbol{B}$ that corresponds to age $x$.

### 1.4.1 TOPALS derivatives

Derivatives of age-specific rates with respect to TOPALS parameters $\boldsymbol{\alpha}$ are $K \times 1$ vectors:

$$\frac{\partial \ln \mu_x}{\partial \boldsymbol{\alpha}} = \boldsymbol{b}_x \quad , \quad \frac{\partial \mu_x}{\partial \boldsymbol{\alpha}} = \mu_x \, \boldsymbol{b}_x$$

For the entire $G \times 1$ vector of group mortality rates, this implies

$$\begin{aligned}
\frac{\partial \boldsymbol{M}'}{\partial \boldsymbol{\alpha}} &= \frac{\partial}{\partial \boldsymbol{\alpha}} \left(\boldsymbol{\mu}' \boldsymbol{W}'\right) \\
&= \left( \frac{\partial \mu_0}{\partial \boldsymbol{\alpha}} \quad \dots \quad \frac{\partial \mu_{A-1}}{\partial \boldsymbol{\alpha}} \right) \boldsymbol{W}' \\
&= \left( \mu_0 \, \boldsymbol{b}_0 \quad \cdots \quad \mu_{A-1} \, \boldsymbol{b}_{A-1} \right) \boldsymbol{W}' \\
&= \boldsymbol{B}' \, diag(\boldsymbol{\mu}) \, \boldsymbol{W}'
\end{aligned}$$

Abbreviate this as

$$\frac{\partial \boldsymbol{M}'}{\partial \boldsymbol{\alpha}} = \boldsymbol{X}'$$

remembering that the $G \times K$ matrix $\boldsymbol{X} = \boldsymbol{W} \, diag(\boldsymbol{\mu}) \, \boldsymbol{B}$ varies with parameters $\boldsymbol{\alpha}$ via the $\mu$ terms in the central diagonal matrix.

[3] Typically $K = 7$, with spline knots fixed at ages $t = 0, 1, 10, 20, 40, 70, 99$. Values in the $k$th column of $B$ are

$$B_{xk} = \begin{cases} \frac{x - t_{k-1}}{t_k - t_{k-1}} & \text{for } x \in [t_{k-1}, t_k] \\ \frac{t_{k+1} - x}{t_{k+1} - t_k} & \text{for } x \in [t_k, t_{k+1}] \\ 0 & \text{otherwise} \end{cases}$$

The $G \times 1$ vector of expected deaths in groups $1 \dots G$ is

$$\hat{\boldsymbol{D}} = \begin{pmatrix} N_1 M_1 \\ \vdots \\ N_G M_G \end{pmatrix} = diag(\boldsymbol{N})\, \boldsymbol{M}$$

and the derivative of its transpose with respect to TOPALS parameters $\boldsymbol{\alpha}$ is a $K \times G$ matrix:

$$\frac{\partial \hat{\boldsymbol{D}}'}{\partial \boldsymbol{\alpha}} = \frac{\partial \boldsymbol{M}'}{\partial \boldsymbol{\alpha}}\, diag(\boldsymbol{N}) \; = \boldsymbol{X}' diag(\boldsymbol{N}) \tag{2}$$

### 1.4.2 Penalized Log Likelihood for TOPALS parameters

With the TOPALS parameterization the log likelihood for a sample $\{D_g, N_g\}$ is

$$\begin{aligned} \ln L(\boldsymbol{\alpha}) &= c + \sum_{g=1}^{G} \Big( D_g \ln M_g(\boldsymbol{\alpha}) - \hat{D}_g(\boldsymbol{\alpha}) \Big) \\ &= c + [\ln M_1 \cdots \ln M_G]\, \boldsymbol{D} - \hat{\boldsymbol{D}}' \mathbf{1} \end{aligned}$$

To stabilize estimates in small populations with few deaths, we add a small penalty term to the log likehood:

$$\begin{aligned} Q(\boldsymbol{\alpha}) &= \ln L(\boldsymbol{\alpha}) - \text{penalty}(\boldsymbol{\alpha}) \\ &= c + [\ln M_1 \cdots \ln M_G]\, \boldsymbol{D} - \hat{\boldsymbol{D}}' \mathbf{1} - \tfrac{1}{2}\, \boldsymbol{\alpha}' \boldsymbol{P} \boldsymbol{\alpha} \end{aligned} \tag{3}$$

where $\boldsymbol{P}$ is a $K \times K$ matrix of constants selected so that the penalty term equals the sum of squared differences between consecutive $\alpha$ parameters – i.e., $\frac{1}{2}\boldsymbol{\alpha}' \boldsymbol{P} \boldsymbol{\alpha} = \sum_{k=2}^{K} (\alpha_k - \alpha_{k-1})^2$. Adding the penalty gives priority to sets of TOPALS parameters $(\alpha_1 \dots \alpha_K)$ that are similar to one another, and thus to log mortality schedules that look more like up-and-down vertical shifts of the standard schedule. For all but the smallest populations the penalty term has virtually no effect on parameter estimates. For very small populations with zero deaths in some age groups, the addition of the penalty stabilizes estimated schedules by borrowing strength across groups.

### 1.4.3 Gradient

We want to select $\boldsymbol{\alpha}$ to maximize $Q$. This requires setting a vector of derivatives equal to zero: $\frac{\partial Q}{\partial \boldsymbol{\alpha}} = 0 \in \mathbb{R}^K$.

Differentiating Eq. (3) with respect to the TOPALS parameters and substituting some of the results above produces an analytical expression for the $K \times 1$ gradient vector of first derivatives:

$$
\begin{aligned}
g(\alpha) = \frac{\partial Q}{\partial \boldsymbol{\alpha}} &= \left[\tfrac{1}{M_1}\tfrac{\partial M_1}{\partial \boldsymbol{\alpha}} \cdots \tfrac{1}{M_G}\tfrac{\partial M_G}{\partial \boldsymbol{\alpha}}\right] \boldsymbol{D} - \frac{\partial \hat{\boldsymbol{D}}'}{\partial \boldsymbol{\alpha}} \mathbf{1} - \boldsymbol{P\alpha} \\
&= \frac{\partial \boldsymbol{M}'}{\partial \boldsymbol{\alpha}} diag(\frac{1}{\boldsymbol{M}})\, \boldsymbol{D} - \frac{\partial \boldsymbol{M}'}{\partial \boldsymbol{\alpha}}\, diag(\boldsymbol{N})\, \mathbf{1} - \boldsymbol{P\alpha} \\
&= \boldsymbol{X}' \left[diag(\frac{1}{\boldsymbol{M}})\, \boldsymbol{D} - N\right] - \boldsymbol{P\alpha}
\end{aligned} \tag{4}
$$

### 1.4.4 Hessian

In Equation (4) the terms that vary with TOPALS coefficients $\alpha$ are $\mathbf{X}$, $diag(\frac{1}{\mathbf{M}})$, and $\alpha$. In order to construct the Hessian matrix of second derivatives, start by considering how the $K \times 1$ gradient changes with a change in only one of the $\alpha$s – say, $\alpha_3$.

By the chain rule,

$$
\begin{aligned}
\frac{\partial}{\partial \alpha_3}\left(\frac{\partial Q}{\partial \boldsymbol{\alpha}}\right) = &\left[\frac{\partial \boldsymbol{X}'}{\partial \alpha_3}\right] \left[diag(\frac{1}{\boldsymbol{M}})\, \boldsymbol{D} - \boldsymbol{N}\right] \\
&+ \boldsymbol{X}' \frac{\partial}{\partial \alpha_3}\left[diag(\frac{1}{\boldsymbol{M}})\right] \boldsymbol{D} \\
&- \frac{\partial}{\partial \alpha_3}\left[\boldsymbol{P\alpha}\right]
\end{aligned} \tag{5}
$$

Equation (5) includes several matrix derivatives:

$$
\begin{aligned}
\underbrace{\left[\frac{\partial \boldsymbol{X}'}{\partial \alpha_3}\right]}_{K \times G} &= \boldsymbol{B}' \left[diag\left(\frac{\partial \mu}{\partial \alpha_3}\right)\right] \boldsymbol{W}' \quad ' \\
&= \boldsymbol{B}' \left[diag\left(e_3' \frac{\partial \mu'}{\partial \alpha}\right)\right] \boldsymbol{W}' \\
&= \boldsymbol{B}' \left[diag\left(b_3' \, diag(\mu)\right)\right] \boldsymbol{W}' \\
&= \boldsymbol{B}' \left[diag\left(b_3 \circ \mu\right)\right] \boldsymbol{W}'
\end{aligned} \tag{6}
$$

where $b_3$ is the $A \times 1$ third column of $\mathbf{B}$ and the $\circ$ symbol represents element-by-element multiplication.

$$
\begin{aligned}
\frac{\partial}{\partial \alpha_3}\left[diag(\frac{1}{\boldsymbol{M}})\right] &= diag\left(\frac{\partial}{\partial \alpha_3}\left[\frac{1}{M_1}\cdots\frac{1}{M_G}\right]\right) \\
&= -diag\left(\left[\frac{\partial M_1}{\partial \alpha_3}{M_1}^{-2}\cdots\frac{\partial M_G}{\partial \alpha_3}{M_G}^{-2}\right]\right) \\
&= -diag\left(\left[\frac{\partial M_1}{\partial \alpha_3}\cdots\frac{\partial M_G}{\partial \alpha_3}\right]\ diag\left(M_1^{-2}\cdots M_G^{-2}\right)\right) \qquad (7) \\
&= -diag\left(\left[b_3'\, diag(\mu)\,\mathbf{W}'\right]\ diag\left(M_1^{-2}\cdots M_G^{-2}\right)\right) \\
&\text{or} \\
&= -diag\left(\frac{(b_3\circ\mu)'w_1}{M_1^2}\cdots\frac{(b_3\circ\mu)'w_G}{M_G^2}\right)
\end{aligned}
$$

where $w_g$ is the $A\times 1$ $g$-th column of $\mathbf{W}$.
Thus the *third* column of the Hessian matrix $Q_{\theta\theta'}$ is

$$
\begin{aligned}
\frac{\partial}{\partial \alpha_3}\left(\frac{\partial Q}{\partial \boldsymbol{\alpha}}\right) &= \boldsymbol{B}'\ \left[diag\left(b_3\circ\mu\right)\right]\boldsymbol{W}'\ \left[diag(\frac{1}{\boldsymbol{M}})\,\boldsymbol{D}-\boldsymbol{N}\right] \\
&\quad -\boldsymbol{X}'\ \left[diag\left(\frac{(b_3\circ\mu)'w_1}{M_1^2}\cdots\frac{(b_3\circ\mu)'w_G}{M_G^2}\right)\right]\boldsymbol{D} \qquad (8) \\
&\quad -[\boldsymbol{P}e_3]
\end{aligned}
$$

and the generic $j$-th column of the Hessian is the $K\times 1$ vector

$$
\begin{aligned}
h_j = \frac{\partial}{\partial \alpha_j}\left(\frac{\partial Q}{\partial \boldsymbol{\alpha}}\right) &= \boldsymbol{B}'\ \left[diag\left(b_j\circ\mu\right)\right]\boldsymbol{W}'\ \left[diag(\frac{1}{\boldsymbol{M}})\,\boldsymbol{D}-\boldsymbol{N}\right] \\
&\quad -\boldsymbol{X}'\ \left[diag\left(\frac{(b_j\circ\mu)'w_1}{M_1^2}\cdots\frac{(b_j\circ\mu)'w_G}{M_G^2}\right)\right]\boldsymbol{D} \qquad (9) \\
&\quad -[\boldsymbol{P}e_j]
\end{aligned}
$$

And the complete $K\times K$ Hessian matrix is the (admittedly complicated, but computable)

$$
\mathbf{H}(\alpha) = \mathbf{Q}_{\theta\theta'} = [h_1\cdots h_K] \qquad (10)
$$

## 1.5 Newton-Raphson Iteration

With analytical expressions for the gradient and Hessian [Equations (4) and (10), respectively] we can now use Newton-Raphson iteration to solve for the $\alpha \in \mathbb{R}^K$ that maximizes the penalized likelihood in Equation (3).

Starting at an arbitrary vector $\alpha_0$ (usually all zeroes) for the TOPALS offsets, we start at $t = 0$ and repeat the following until convergence:

1. calculate the gradient and Hessian at the current parameter values: $g(\alpha_t)$ and $H(\alpha_t)$
2. using the (multivariate) quadratic approximation to the objective function, generate a new parameter vector at which the gradient should approximately equal zero:

$$\alpha_{t+1} = \alpha_t - [H(\alpha_t)]^{-1} g(\alpha_t)$$

3. increment $t$ by one and return to step 1