1 Extended Appendix

1.1 Newton-Raphson

In **??**, we discussed the general MLE projection search function. Newton-Raphson is another popular way of choosing a direction $\mathbf{d}^{(i)}$. In this particular maximization problem, $\alpha^{(i)} \cdot \mathbf{d}^{(i)}$ is equivalent to $\boldsymbol{\mathcal{J}}_n^{-1}\left(\boldsymbol{\theta}^{(i)}\right) \cdot \nabla\ell\left(\boldsymbol{\theta}^{(i)}\right)$, where $\boldsymbol{\mathcal{J}}_n^{-1}\left(\boldsymbol{\theta}^{(i)}\right)$ denotes the inverse (or pseudoinverse) of the observed information matrix.

This particular choice derives from the fact that it is generalization of Newton's method to higher dimensions. Consider the following proof.

Proof. At a local maximum, the gradient of $\ell(\boldsymbol{\theta} \mid \mathbb{F}_n)$ is zero,

$$\nabla\ell(\boldsymbol{\theta} \mid \mathbb{F}_n) = \mathbf{0}.$$

Solutions satisfying this system of equations are known as the stationary points of the function ℓ. To solve this system of equations, we first solve an easier linear system of equations that approximates $\nabla\ell(\boldsymbol{\theta} \mid \mathbb{F}_n) = \mathbf{0}$. Let the current solution be $\boldsymbol{\theta}^i$. Let $\mathrm{g}(\boldsymbol{\theta})$ be the linear approximation of $\nabla\ell(\boldsymbol{\theta} \mid \mathbb{F}_n)$ at $\boldsymbol{\theta}^i$,

$$\begin{aligned}
\mathrm{g}(\boldsymbol{\theta} \mid \mathbb{F}_n) &= \nabla^2\ell\left(\boldsymbol{\theta}^{(i)} \mid \mathbb{F}_n\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}\right) + \nabla\ell\left(\boldsymbol{\theta}^{(i)} \mid \mathbb{F}_n\right) \\
&= \mathcal{H}\left(\ell\left(\boldsymbol{\theta}^{(i)} \mid \mathbb{F}_n\right)\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}\right) + \nabla\ell\left(\boldsymbol{\theta}^{(i)} \mid \mathbb{F}_n\right),
\end{aligned}$$

which is a reasonable approximation of $\ell(\boldsymbol{\theta} \mid \mathbb{F}_n)$ if $\boldsymbol{\theta} \approx \boldsymbol{\theta}^{(i)}$. Solving the roots of this linear equation,

$$\begin{aligned}
\mathbf{0} &= \mathcal{H}\left(\ell\left(\boldsymbol{\theta}^{(i)}\right)\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}\right) + \nabla\ell\left(\boldsymbol{\theta}^{(i)}\right) \\
\mathcal{H}\left(\ell\left(\boldsymbol{\theta}^{(i)}\right)\right) \cdot \boldsymbol{\theta} &= \mathcal{H}\left(\ell\left(\boldsymbol{\theta}^{(i)}\right)\right) \cdot \boldsymbol{\theta}^{(i)} - \nabla\ell\left(\boldsymbol{\theta}^{(i)}\right) \\
\mathcal{H}^{-1}\left(\ell\left(\boldsymbol{\theta}^{(i)}\right)\right)\mathcal{H}\left(\ell\left(\boldsymbol{\theta}^{(i)}\right)\right) \cdot \boldsymbol{\theta} &= \mathcal{H}^{-1}\left(\ell\left(\boldsymbol{\theta}^{(i)}\right)\right)\left(\mathcal{H}\left(\ell\left(\boldsymbol{\theta}^{(i)}\right)\right) \cdot \boldsymbol{\theta}^{(i)} - \nabla\ell\left(\boldsymbol{\theta}^{(i)}\right)\right) \\
\boldsymbol{\theta} &= \boldsymbol{\theta}^{(i)} - \mathcal{H}^{-1}\left(\ell\left(\boldsymbol{\theta}^{(i)}\right)\right)\nabla\ell\left(\boldsymbol{\theta}^{(i)}\right)
\end{aligned}$$

Notice that $\mathcal{H}^{-1}\left(\ell\left(\boldsymbol{\theta}^{(i)}\right)\right)$ is the observed information matrix evaluated at $\boldsymbol{\theta}^{(i)}$. Performing this substitution we arrive at

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{(i)} + \boldsymbol{\mathcal{J}}_n^{-1}\left(\boldsymbol{\theta}^{(i)}\right)\nabla\ell\left(\boldsymbol{\theta}^{(i)}\right)$$

where $\boldsymbol{\mathcal{J}}_n$ is the observed information matrix and $\nabla\ell$ is the gradient of ℓ with respect to $\boldsymbol{\theta}$. Thus, the direction $\alpha^{(i)} \cdot \mathbf{d}^{(i)}$ is a function of the gradient and the Hessian. □

1.2 Arbitrary constraints

If one or more of the constraints are non-linear or non-convex then **??** is not an appropriate choice.

Without loss of generality, let the constraints be defined in terms of k inequalities, $\mathrm{c_i}(\boldsymbol{\theta}) \geq 0$ for $i = 1, \ldots, k$. For example, if the i^{th} constraint is $\theta_1 \cdot \theta_2 \leq \theta_3$, then $\mathrm{c_i}(\boldsymbol{\theta}) = \theta_3 - \theta_1 \cdot \theta_2 \geq 0$. There are many ways to try to solve these kind of problems, although they are more challenging.

One approach, known as the *Penalty method*, replaces the d-constrained maximization problem,

$$\begin{aligned} \hat{\boldsymbol{\theta}} = \arg\max \quad & \ell(\boldsymbol{\theta}) \\ \text{subject to} \quad & \mathrm{c_i}(\boldsymbol{\theta}) \geq 0, i = 1, \ldots, d \end{aligned}$$

with a sequence of unconstrained maximization problems

$$\hat{\boldsymbol{\theta}}_k = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) + \sigma_k \sum_{i=1}^{d} \min\left(0, \mathrm{c_i}(\boldsymbol{\theta})\right)^2, k = 1, \ldots$$

such that infeasible solutions are progressively penalized as k increases. It has been proven that the sequence of solutions to the unconstrained maximization problem converges to the solution to the constrained maximization problem, i.e., $\lim_{k \to \infty} \hat{\boldsymbol{\theta}}_k \to \hat{\boldsymbol{\theta}}$, although this convergence can be extremely slow.

Specifically, we solve $\hat{\boldsymbol{\theta}}_k$ for $k = 1, \ldots$ using some numerical solver, such as **??**, and we let $\hat{\boldsymbol{\theta}}_k$ be the initial starting value for the $(k+1)$-th unconstrained maximization problem. In addition, at the start, the penalty factor is some small value, e.g., $\sigma_1 = \epsilon$, and we increase it by some constant factor after each iteration, $\sigma_{p+1} = \beta \cdot \sigma_p, \beta > 1$. Thus, in the early rounds searches outside of the feasible parameter space are barely penalized and so can explore unfettered and in the late rounds searches outside of the feasible parameter space are severely penalized and so are pushed to explore the feasible parameter space.

1.3 Conditional probability of failure given uncertain failure time

Theorem. *The probability $\mathbb{P}_{\boldsymbol{\theta}^\star}(k \mid t_a, t_b)$ that a system failure that occurred at some time $t \in (t_a, t_b)$ (with uniform probability) was caused by component k is*

$$\mathbb{P}_{\boldsymbol{\theta}^\star}(k \mid t_a, t_b) = \left[\int_{t_a}^{t_b} f_k(t \mid \boldsymbol{\theta}_k^\star) \prod_{\substack{p=1 \\ p \neq j}}^{m} R_p(t \mid \boldsymbol{\theta}_p^\star) \mathrm{d}t\right] \left[R(t_a \mid \boldsymbol{\theta}^\star) - R(t_b \mid \boldsymbol{\theta}^\star)\right]^{-1}.$$

Proof. Let it be given that the system failure occurred at some time t in the interval (t_a, t_b). Observe that component k caused the system failure if it failed at time $t \in (t_a, t_b)$ and the other components survived longer than component k. Thus, the conditional probability that component k caused the system failure given a system failure at time $t \in (t_a, t_b)$ is

$$\Pr\left[t_a < T_k < t_b \cap_{j \neq k} T_p > T_k \mid t_a < T < t_b\right],$$

where T_j is a random variable denoting the lifetime of component j for $j = 1, \ldots, m$ and T is a random variable denoting the lifetime of a series system composed of those m components.

By Bayes law, the above conditional probability is equivalent to

$$\frac{\Pr\left[t_a < T_k < t_b \cap_{p \neq k} T_p > T_k\right]}{\Pr\left[t_a < T < t_b\right]} = \frac{\int_{t_a}^{t_b} f_k(t) \prod_{p \neq k} R_p(t) \mathrm{d}t}{R(t_a) - R(t_b)}$$

□

1.4 Calculating confidence intervals using entropy minimization and maximization

Theorem. *The $(1-\alpha) \cdot 100\%$-confidence interval for parameter $\mathbb{E}[N_{t,\sigma}]$ is (L_t, U_t)*

$$L_t = \sum_{n=1}^{m} n \mathbb{P}_{\boldsymbol{\theta}_L}(\sigma(n) \mid t)$$
$$U_t = \sum_{n=1}^{m} n \mathbb{P}_{\boldsymbol{\theta}_H}(\sigma(n) \mid t)$$

where

$$\boldsymbol{\theta}_L = \underset{\boldsymbol{\theta} \in D}{\arg\min} \sum_{j=1}^{m} \mathbb{P}_{\boldsymbol{\theta}}(j \mid t) \log \mathbb{P}_{\boldsymbol{\theta}}(j \mid t).$$

and

$$\boldsymbol{\theta}_U = \underset{\boldsymbol{\theta} \in D}{\arg\max} \sum_{j=1}^{m} \mathbb{P}_{\boldsymbol{\theta}}(j \mid t) \log \mathbb{P}_{\boldsymbol{\theta}}(j \mid t).$$

where D contains the q confidence intervals for the q parameters.

1.5 Observed Fisher information

If no expectation is taken in **??**, we have instead the observed Fisher information matrix $\boldsymbol{\mathcal{J}}$.

Definition. *Observed Fisher information matrix. Consider the log-likelihood function* $\ell(\boldsymbol{\theta} \in \boldsymbol{\Theta} \mid \mathbb{F}_n)$ *where* $\boldsymbol{\Theta} \subset \mathbb{R}^q$*. The observed Fisher information matrix* $\boldsymbol{\mathcal{J}}_n(\boldsymbol{\theta})$ *for an* $\mathbb{F}_n$ *sample of size* n *is a* q*-by-*q *symmetric matrix whose* (i, j)*-th element is the negative of the second partial derivative of the log-likelihood* $\ell(\boldsymbol{\theta} \mid \mathbb{F}_n)$ *with respect to parameters* θ_i *and* θ_j*. That is,*

$$\boldsymbol{\mathcal{J}}_n(\boldsymbol{\theta}) = -\,\mathcal{H}\left(\ell(\boldsymbol{\theta} \mid \mathbb{F}_n)\right)$$

where $\mathcal{H}$ *is the Hessian. The hessian of function* f *with respect to vector parameter* $\boldsymbol{\theta} \subset \mathbb{R}^q$ *is*

$$\mathcal{H}\left(f(\boldsymbol{\theta})\right) = -\begin{bmatrix} \frac{\partial^2 f(\boldsymbol{\theta})}{\partial\theta_1^2} & \frac{\partial^2 f(\boldsymbol{\theta})}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 f(\boldsymbol{\theta})}{\partial\theta_1\partial\theta_q} \\ \frac{\partial^2 f(\boldsymbol{\theta})}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 f(\boldsymbol{\theta})}{\partial\theta_2^2} & \cdots & \frac{\partial^2 f(\boldsymbol{\theta})}{\partial\theta_2\partial\theta_q} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 f(\boldsymbol{\theta})}{\partial\theta_q\partial\theta_1} & \frac{\partial^2 f(\boldsymbol{\theta})}{\partial\theta_q\partial\theta_2} & \cdots & \frac{\partial^2 f(\boldsymbol{\theta})}{\partial\theta_q^2} \end{bmatrix}.$$

The hessian $\mathcal{H}$ and observed information matrix $\boldsymbol{\mathcal{J}}$ are used in **??**.