Empirical Fisher, Gradient Covariance, Gauss-Newton Matrix, Fisher and Hessian

Jingfeng Wu

Created: July 2019
Last updated: June 27, 2020

This short notes aim to clarify the relationship between the so called *Empirical Fisher* $\tilde{F}$, *Gradient Covariance* C, *Gauss-Newton matrix* G, *Fisher* F and *Hessian* H. See [1, 3, 4] for more detailed discussions.

1 Definitions

Let $p_D(y|x)$ be the population distribution, and $p_\theta(y|x)$ be our model parameterized by θ. $D_N := \{x_n, y_n\}_{n=1}^N$ denote the observed data pair.

To learn the model from observations, consider the max log-likelihood estimation, i.e., $\ell(x, y; \theta) = -\log p_\theta(y|x)$ we minimize the following loss:

$$\min_\theta L(\theta) := -\mathbb{E}_{D_N} \log p_\theta(y_n|x_n) = \mathbb{E}_{D_N} \ell(x_n, y_n; \theta).$$

Let $g_\theta(x, y) := \nabla_\theta \ell(x, y; \theta)$ be the gradient of loss at point (x, y).

Definition 1 (Empirical Fisher)**.**

$$\begin{aligned}\tilde{F}(\theta) :=& \mathbb{E}_{(x_n,y_n)\in D_N} \left[\nabla_\theta \log p_\theta(y_n|x_n) \nabla_\theta \log p_\theta(y_n|x_n)^T\right] \\ =& \mathbb{E}_{(x_n,y_n)\in D_N} \left[g_\theta(x_n, y_n) g_\theta(x_n, y_n)^T\right].\end{aligned}$$

Definition 2 (Fisher)**.**

$$\begin{aligned}F(\theta) :=& \mathbb{E}_{x\in D_N} \mathbb{E}_{y\sim p_\theta(y|x_n)} \left[\nabla_\theta \log p_\theta(y|x_n) \nabla_\theta \log p_\theta(y|x_n)^T\right] \\ =& \mathbb{E}_{x_n\in D_N} \mathbb{E}_{y\sim p_\theta(y|x_n)} \left[g_\theta(x_n, y) g_\theta(x_n, y)^T\right].\end{aligned}$$

Definition 3 (Gradient Covariance)**.**

$$\begin{aligned}C(\theta) :=& \mathbb{E}_{(x_n,y_n)\in D_N} \left[\left(g_\theta(x_n, y_n) - \nabla_\theta L(\theta)\right) \left(g_\theta(x_n, y_n) - \nabla_\theta L(\theta)\right)^T\right] \\ =& \tilde{F}(\theta) - \nabla_\theta L(\theta) \nabla_\theta L(\theta)^T.\end{aligned}$$

Definition 4 (Hessian)**.**

$$H(\theta) := \nabla_\theta^2 L(\theta).$$

2 Clarification

1. All the notations are actually "empirical", i.e., none of them rely on the population distribution of observed data.

2. Fisher is originally adopted for scoring (Fisher's scoring [2]). Fisher involves the expectation of label y w.r.t. current learned model $p_\theta(y|x)$, which measures the information contained in the model, thus it is also called the *Fisher information*.

3. The convention of Empirical Fisher is a little bit misleading. Empirical Fisher is not the empirical realization of Fisher. As pointed before, the Fisher itself is an empirical notation. In contrast, Empirical Fisher should be called *the second moment of gradients*, as $C(\theta) = \tilde{F}(\theta) - \nabla_\theta L(\theta) \nabla_\theta L(\theta)^T$.

Now we are crystal clear about the exact definition of each matrices. However, in literature, the usage of the matrices is often miss specified, especially for Fisher and Empirical Fisher. Lots of papers mistakenly use Empirical Fisher as Fisher. Indeed, Empirical Fisher approximates Fisher under some conditions, which are described below.

3 The approximation of the matrices

Now we talk about when the matrices would approximate to the others.

Proposition 1 ($C \approx \tilde{F}$). *For any minima θ^* of loss $L(\theta)$, local or global, true or not true, we have*

$$C(\theta^*) = \tilde{F}(\theta^*).$$

Proposition 2 ($\tilde{F} \approx F$). *If 1) the number of observed data is large enough $N \to \infty$, then 2) for the true minima θ_0 (i.e., $p_{\theta_0}(y|x) = p_D(y|x)$), we have*

$$\tilde{F}(\theta_0) \approx F(\theta_0).$$

Proposition 3 ($F \approx H$). *If 1) the number of observed data is large enough $N \to \infty$, then 2) for the true minima θ_0 (i.e., $p_{\theta_0}(y|x) = p_D(y|x)$), we have*

$$F(\theta_0) \approx \tilde{H}(\theta_0).$$

4 The Generalized Gauss-Newton Decomposition

We consider the negative log likelihood loss[1]:

$$\begin{aligned}
&\ell(x;\theta) = -\log p(x;\theta), \\
&L(\theta) = \mathbb{E}_x \ell(x;\theta) = -\mathbb{E}_x \log p(x;\theta).
\end{aligned}$$

The following Gauss-Newton decomposition connects the Hessian $H(\theta) := \nabla^2_\theta L(\theta)$ and the Fisher $F(\theta)$:

$$\begin{aligned}
H(\theta) &= -\mathbb{E}_x \nabla_\theta \left(\frac{\partial l}{\partial p} \frac{\partial p(x;\theta)}{\partial \theta} \right) \\
&= -\mathbb{E}_x \nabla_\theta \left(\frac{1}{p(x;\theta)} \frac{\partial p(x;\theta)}{\partial \theta} \right) \\
&= \mathbb{E}_x \left(\frac{1}{p(x;\theta)^2} \frac{\partial p(x;\theta)}{\partial \theta}^T \frac{\partial p(x;\theta)}{\partial \theta} \right) - \mathbb{E}_x \left(\frac{1}{p(x;\theta)} \frac{\partial^2 p(x;\theta)}{\partial \theta^2} \right) \\
&= \mathbb{E}_x \left(\frac{\partial \log p(x;\theta)}{\partial \theta}^T \frac{\partial \log p(x;\theta)}{\partial \theta} \right) - \mathbb{E}_x \left(\frac{1}{p(x;\theta)} \frac{\partial^2 p(x;\theta)}{\partial \theta^2} \right) \\
&= F(\theta) - \mathbb{E}_x \left(\frac{1}{p(x;\theta)} \frac{\partial^2 p(x;\theta)}{\partial \theta^2} \right).
\end{aligned}$$

References

[1] Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation, 2019.

[2] Nicholas T Longford. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74(4):817–827, 1987.

[3] James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.

[4] Valentin Thomas, Fabian Pedregosa, Bart van Merriënboer, Pierre-Antoine Mangazol, Yoshua Bengio, and Nicolas Le Roux. Information matrices and generalization, 2019.

[1] For L_2 loss, there is a similar form of decomposition, too.