Misspecified Likelihood and Misspecification Testing

Jesper Riis-Vestergaard Sørensen

December 3, 2025

Abstract

Our discussion of maximum likelihood has thus far presumed correct model specification. We here introduce the possibility of model misspecification and discuss the consequences thereof. We finally present the White (1982) Information Matrix (IM) Test, which can be viewed as a test of correct specification.

1 Framework: Likelihood

- We are interested in the conditional distribution $\mathrm{D}[\boldsymbol{Y}_i|\boldsymbol{X}_i]$ (or features thereof) of $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i$, where $\boldsymbol{Y}_i$ has support $\mathcal{Y} \subseteq \mathbb{R}^{d_Y}$, $\boldsymbol{X}_i$ has support $\mathcal{X} \subseteq \mathbb{R}^{d_X}$, and the dimensions d_Y and d_X are fixed and finite.

- Corresponding to this (conditional) distribution there is a true (conditional) density $p_o : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}_+$ function. This density could represent one or more discrete random variables (having probability mass), continuous random variables (having probability density), or a mix thereof. Hence, "density" is understood in a broad sense.

- Our (parametric) model for p_o consists of a family of candidates for p_o,

$$\mathcal{P} := \{\mathcal{Y} \times \mathcal{X} \ni (\boldsymbol{y}, \boldsymbol{x}) \mapsto p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) | \boldsymbol{\theta} \in \boldsymbol{\Theta}\},$$

  parameterized by vectors $\boldsymbol{\theta}$ from a space $\Theta \subseteq \mathbb{R}^d$ of fixed and finite dimension d.

- Our discussion of maximum likelihood has thus far presumed that the candidates $\mathcal{P}$ are *legitimate densities*, meaning that they are (i) *nonnegative*

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \geqslant 0 \text{ for all } (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}) \in \mathcal{Y} \times \mathcal{X} \times \Theta;$$

and, (ii) *integrate to one* (against ν)

$$\int_{\mathcal{Y}} p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\nu(\mathrm{d}\boldsymbol{y}) = 1 \text{ for all } (\boldsymbol{x},\boldsymbol{\theta}) \in \mathcal{X} \times \Theta.$$

Technical corner: The ν notation is there to capture different types of random variables.

- If $\boldsymbol{Y}_i$ is *discrete*, then p_o is a probability mass function (PMF). In this case, integration against ν means summing through the (then) discrete set $\mathcal{Y}$,

$$\int_{\mathcal{Y}} p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\nu(\mathrm{d}\boldsymbol{y}) = \sum_{\boldsymbol{y}\in\mathcal{Y}} p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}),$$

  and legitimacy means that the candidates correspond to probabilities.

- If $\boldsymbol{Y}_i$ is (absolutely) *continuous*, then p_o is a probability density function (PDF). In this case, integration against ν is just "ordinary" integration,

$$\int_{\mathcal{Y}} p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\nu(\mathrm{d}\boldsymbol{y}) = \int_{\mathcal{Y}} p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\mathrm{d}\boldsymbol{y}.$$

- Moreover, we have presumed that $\mathcal{P}$ is *correctly specified.* This assumption means that at least one of our candidate densities recovers the true one, i.e., there is a candidate parameter $\boldsymbol{\theta} \in \Theta$ such that $p_o(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$ for all $(\boldsymbol{y},\boldsymbol{x}) \in \mathcal{Y} \times \mathcal{X}$. Visually, for such a $\boldsymbol{\theta}$, the functions p_o and $p(\cdot|\cdot,\boldsymbol{\theta})$ have identical graphs, so that one cannot tell them apart. We can therefore think of the data as if it stems from a distribution based on the density $p(\cdot|\cdot,\boldsymbol{\theta})$. We therefore refer to such a candidate parameter as the *true theta,* often denoted $\boldsymbol{\theta}_o$.

- Presuming correct specification, the question of *identification* is then: Can we back out the true theta?

- Correct specification means that the set $\mathcal{P}$ of candidate densities is large enough so as to include the true density, $p_o \in \mathcal{P}$. In practice, one could have formulated too narrow in model, missing out on certain features of the data. What does maximum likelihood then deliver? And one can one device a test for correct specification?

2 Model Misspecification and Pseudo Truth

- We next maintain legitimate candidates densities but allow for the possibility that $p_o \notin \mathcal{P}$. Stated in terms of model parameters, there could be no $\boldsymbol{\theta} \in \Theta$ for which

$p_o(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ for all $(\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{Y} \times \mathcal{X}$.

- We have shown that, presuming identification (which requires correct specification), the true theta $\boldsymbol{\theta}_o$ can be viewed as the unique solution to the population problem (PP) in which one maximizes the expected model log-likelihood,

$$\boldsymbol{\theta}_o = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathrm{E}[\ln p(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\theta})].$$

- When $p_o \notin \mathcal{P}$, the model becomes an *approximation*, and one cannot speak of a "true theta."

- There could, however, still be a solution to the above problem. Supposing not only that a solution exists but also that it is unique, denote the unique solution

$$\boldsymbol{\theta}_* := \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathrm{E}[\ln p(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\theta})].$$

- The parameter $\boldsymbol{\theta}_*$ yields the "best" fitting model in a sense to be made precise. Since shifting the objective function up/down by a constant has no impact on the solution, maximizing $\boldsymbol{\theta} \mapsto \ln p(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\theta})$ is equivalent to maximizing $\boldsymbol{\theta} \mapsto \mathrm{E}[\ln p(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\theta})] - \mathrm{E}[\ln p_o(\boldsymbol{Y}_i|\boldsymbol{X}_i)]$. Hence

$$\begin{aligned}
\boldsymbol{\theta}_* &= \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathrm{E}\left[\ln\left(\frac{p(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\theta})}{p_o(\boldsymbol{Y}_i|\boldsymbol{X}_i)}\right)\right] && \text{(shift)} \\
&= \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathrm{E}\left[\int_{\mathcal{Y}} \ln\left(\frac{p(\boldsymbol{y}|\boldsymbol{X}_i, \boldsymbol{\theta})}{p_o(\boldsymbol{y}|\boldsymbol{X}_i)}\right) p_o(\boldsymbol{y}|\boldsymbol{X}_i)\nu(\mathrm{d}\boldsymbol{y})\right] && \text{(iterate)} \\
&= \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathrm{E}\left[\int_{\mathcal{Y}} \ln\left(\frac{p_o(\boldsymbol{y}|\boldsymbol{X}_i)}{p(\boldsymbol{y}|\boldsymbol{X}_i, \boldsymbol{\theta})}\right) p_o(\boldsymbol{y}|\boldsymbol{X}_i)\nu(\mathrm{d}\boldsymbol{y})\right] && \text{(sign flip)}
\end{aligned}$$

  where the outer expectation is over the distribution of $\boldsymbol{X}_i$.

- The inner expectation is the (conditional on $\boldsymbol{X}_i$) *Kullback–Leibler* (KL) *divergence* (also known as the *relative entropy*) from $p(\cdot|\boldsymbol{X}_i, \boldsymbol{\theta})$ to $p_o(\cdot|\boldsymbol{X}_i)$. It is a measure of distance between the two densities/the distributions they represent. The parameter $\boldsymbol{\theta}_*$ minimizes this distance, thus yielding the best fitting density $p(\cdot|\boldsymbol{X}_i, \boldsymbol{\theta}_*)$ under the KL notion of distance.

- When the model is correctly specified (and identified), $\boldsymbol{\theta}_*$ reduces to the true theta $\boldsymbol{\theta}_o$. For this reason, $\boldsymbol{\theta}_*$ is often referred to as the *pseudo-true theta*.

3 Misspecificed Maximum Likelihood

- Assuming access to iid observations $\{(\boldsymbol{Y}_i, \boldsymbol{X}_i)\}_{i=1}^n$, we can still consider fitting the model $\mathcal{P}$ using maximum likelihood. A maximum likelihood estimator (MLE)

$$\widehat{\boldsymbol{\theta}}_n \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell_i(\boldsymbol{\theta}) \right\}, \quad \ell_i(\boldsymbol{\theta}) := \ln p(\boldsymbol{Y}_i | \boldsymbol{X}_i, \boldsymbol{\theta}),$$

  then becomes an estimator of the pseudo-true parameter $\boldsymbol{\theta}_*$. (Sometimes the names *pseudo-MLE* or *quasi-MLE* are used to stress that the model could be misspecified.)

- As long as our model is "sufficiently nice" (think: likelihoods stay away from 0/1 and are twice differentiable), we can still use a Taylor expansion and derive asymptotic normality of the shifted (by $\boldsymbol{\theta}_*$) and scaled MLE

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*\big) \xrightarrow{D} \mathrm{N}\big(\mathbf{0}_{d\times 1}, \boldsymbol{A}_*^{-1}\boldsymbol{B}_*\boldsymbol{A}_*^{-1}\big),$$

  where the $d \times d$ limit variance is of the sandwich form $\boldsymbol{A}_*^{-1}\boldsymbol{B}_*\boldsymbol{A}_*^{-1}$, with

$$\boldsymbol{A}_* := -\mathrm{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell_i(\boldsymbol{\theta}_*)\right]$$

  being the expected Hessian of the (negative) log-likelihood contribution, and

$$\boldsymbol{B}_* := \mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_*)\frac{\partial}{\partial\boldsymbol{\theta}^\top}\ell_i(\boldsymbol{\theta}_*)\right]$$

  the expected outer product of the score, both of which are evaluated at the pseudo-true theta.

- If the model is correctly specified, then $\boldsymbol{A}_* = \boldsymbol{B}_*$ per the information matrix equalities. However, without the guarantee of correct specification, the variance sandwich does not simplify.

- We can still estimate the asymptotic variance $\mathrm{Avar}(\widehat{\boldsymbol{\theta}}_n) := \boldsymbol{A}_*^{-1}\boldsymbol{B}_*\boldsymbol{A}_*^{-1}/n$ of the MLE, and use it for inference purposes. But that inference then concerns the pseudo-true parameter.

4 Misspecification Testing: White's IM Test

- Under correct specification (and a "sufficiently nice" model), the (unconditional) information matrix equality states that

$$-\mathrm{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell_i(\boldsymbol{\theta}_o)\right] = \mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_o)\frac{\partial}{\partial\boldsymbol{\theta}^\top}\ell_i(\boldsymbol{\theta}_o)\right].$$

  Correct specification (and identification) therefore implies that

$$-\mathrm{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell_i(\boldsymbol{\theta}_*)\right] = \mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_*)\frac{\partial}{\partial\boldsymbol{\theta}^\top}\ell_i(\boldsymbol{\theta}_*)\right],$$

  where we have swapped the true theta for the pseudo-true one.

- Take the latter display as our *null hypothesis* H_0 to be tested. The null is pitted against the alternative H_1 that the above (matrix) equality fails. A rejection of H_0 is interpreted as a rejection of correct specification, meaning that the model is misspecified. This observation is the basis of Hal *White's Information Matrix* (IM) *Test.*

- The matrices $\boldsymbol{A}_*$ and $\boldsymbol{B}_*$ are symmetric, so that we only need to compare their upper triangular parts. To this end, abbreviate $\boldsymbol{w} := (\boldsymbol{y}, \boldsymbol{x})$ and define *d*iscrepancy functions $d_\ell : \mathcal{Y} \times \mathcal{X} \times \Theta \to \mathbb{R}, \ell = 1, 2, \ldots, q$, where $q := d(d+1)/2$, by

$$d_\ell(\boldsymbol{w}, \boldsymbol{\theta}) := \frac{\partial}{\partial\theta_j}\ln p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \cdot \frac{\partial}{\partial\theta_k}\ln p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) - \frac{\partial^2}{\partial\theta_j\partial\theta_k}\ln p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$$

  for $j = 1, 2, \ldots, d$ and $k = j, j+1, \ldots, d$. When evaluated at $\boldsymbol{W}_i$ and $\boldsymbol{\theta}_*$, the null hypothesis can then be written as

$$\mathrm{E}[d_\ell(\boldsymbol{W}_i, \boldsymbol{\theta}_*)] = 0 \text{ for all } \ell = 1, 2, \ldots, q.$$

  Technical corner: Some d_ℓ may be identically zero as a consequence of the model, meaning that no discrepancy can arise from these elements. As such d_ℓ should be dropped from consideration, in what follows q is to be interpreted as the number of *nonredundant* discrepancy functions (which have been suitably relabelled).

- Define $\boldsymbol{D}_n : \Theta \to \mathbb{R}^q$ by

$$\boldsymbol{D}_n(\boldsymbol{\theta}) := \begin{bmatrix} n^{-1}\sum_{i=1}^n d_1(\boldsymbol{W}_i, \boldsymbol{\theta}) \\ n^{-1}\sum_{i=1}^n d_2(\boldsymbol{W}_i, \boldsymbol{\theta}) \\ \vdots \\ n^{-1}\sum_{i=1}^n d_q(\boldsymbol{W}_i, \boldsymbol{\theta}) \end{bmatrix}. \qquad (q \times 1)$$

  The test will be based on $\boldsymbol{D}_n(\widehat{\boldsymbol{\theta}}_n)$, which (roughly speaking) are the deviations from IM equality.

- One can show that $\sqrt{n}\boldsymbol{D}_n(\widehat{\boldsymbol{\theta}}_n)$ is asymptotically distributed as $\mathrm{N}(\mathbf{0}_{q\times 1}, \boldsymbol{V}_*)$. Moreover, assuming that these discrepancy functions are differentiable (i.e., an even nicer model), its limit variance can be consistently estimated. To this end, define the Jacobian mapping $\nabla\boldsymbol{D}_n : \Theta \to \mathbb{R}^{q\times d}$ by

$$\nabla\boldsymbol{D}_n(\boldsymbol{\theta}) = \begin{bmatrix} n^{-1}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}} d_1(\boldsymbol{W}_i, \boldsymbol{\theta}) \\ n^{-1}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}} d_2(\boldsymbol{W}_i, \boldsymbol{\theta}) \\ \vdots \\ n^{-1}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}} d_q(\boldsymbol{W}_i, \boldsymbol{\theta}) \end{bmatrix},$$

  and the Hessian mapping $\boldsymbol{A}_n : \Theta \to \mathbb{R}^{d\times d}$ by

$$\boldsymbol{A}_n(\boldsymbol{\theta}) := -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell_i(\boldsymbol{\theta}).$$

  Then the variance estimator is

$$\widehat{\boldsymbol{V}}_n := \frac{1}{n}\sum_{i=1}^n \Big[\underbrace{\boldsymbol{d}(\boldsymbol{W}_i, \widehat{\boldsymbol{\theta}}_n)}_{q\times 1} + \underbrace{\nabla\boldsymbol{D}_n(\widehat{\boldsymbol{\theta}}_n)}_{q\times d}\underbrace{\boldsymbol{A}_n(\widehat{\boldsymbol{\theta}}_n)}_{d\times d}\underbrace{\frac{\partial}{\partial\boldsymbol{\theta}}\ell_i(\widehat{\boldsymbol{\theta}}_n)}_{d\times 1}\Big]$$
$$\cdot\Big[\boldsymbol{d}(\boldsymbol{W}_i, \widehat{\boldsymbol{\theta}}_n) + \nabla\boldsymbol{D}_n(\widehat{\boldsymbol{\theta}}_n)\boldsymbol{A}_n(\widehat{\boldsymbol{\theta}}_n)\frac{\partial}{\partial\boldsymbol{\theta}}\ell_i(\widehat{\boldsymbol{\theta}}_n)\Big]^\top.$$

- The *IM test statistic* is

$$\mathrm{IM}_n := n\boldsymbol{D}_n(\widehat{\boldsymbol{\theta}}_n)^\top\widehat{\boldsymbol{V}}_n^{-1}\boldsymbol{D}_n(\widehat{\boldsymbol{\theta}}_n).$$

  which under the null is asymptotically distributed as chi-square, $\mathrm{IM}_n \to_D \chi^2_q$ as $n \to \infty$.

- For a given significance level $\alpha \in (0,1)$, the *IM test* rejects H_0 if and only if IM_n exceeds the $(1-\alpha)$-quantile of χ^2_q.

References

White, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, 1–25. [1]