15 Model Selection

15.1 KLIC

Suppose a random sample is $\mathbf{y} = y_1, , ..., y_n$ has unknown density $\mathbf{f}(\mathbf{y}) = \prod f(y_i)$.

A model density $\mathbf{g}(\mathbf{y}) = \prod g(y_i)$.

How can we assess the "fit" of $\mathbf{g}$ as an approximation to $\mathbf{f}$?

One useful measure is the Kullback-Leibler information criterion (KLIC)

$$KLIC(\mathbf{f}, \mathbf{g}) = \int \mathbf{f}(\mathbf{y}) \log \left(\frac{\mathbf{f}(\mathbf{y})}{\mathbf{g}(\mathbf{y})} \right) d\mathbf{y}$$

You can decompose the KLIC as

$$\begin{aligned} KLIC(\mathbf{f}, \mathbf{g}) &= \int \mathbf{f}(\mathbf{y}) \log \mathbf{f}(\mathbf{y}) d\mathbf{y} - \int \mathbf{f}(\mathbf{y}) \log \mathbf{g}(\mathbf{y}) d\mathbf{y} \\ &= C_f - E \log \mathbf{g}(\mathbf{y}) \end{aligned}$$

The constant $C_f = \int \mathbf{f}(\mathbf{y}) \log \mathbf{f}(\mathbf{y}) d\mathbf{y}$ is independent of the model g.

Notice that $KLIC(f, g) \geq 0$, and $KLIC(f, g) = 0$ iff $g = f$. Thus a "good" approximating model g is one with a low KLIC.

15.2 Estimation

Let the model density $g(y, \theta)$ depend on a parameter vector θ. The negative log-likelihood function is

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} \log g(y_i, \theta) = -\log \mathbf{g}(\mathbf{y}, \theta)$$

and the MLE is $\hat{\theta} = \text{argmin}_\theta \mathcal{L}(\theta)$. Sometimes this is called a "quasi-MLE" when $g(y, \theta)$ is acknowledged to be an approximation, rather than the truth.

Let the minimizer of $-E \log g(y, \theta)$ be written θ_0 and called the pseudo-true value. This value also minimizes $KLIC(f, g(\theta))$. As the likelihood divided by n is an estimator of $-E \log g(y, \theta)$, the MLE $\hat{\theta}$ converges in probability to θ_0. That is,

$$\hat{\theta} \to_p \theta_0 = \underset{\theta}{\text{argmin}}\, KLIC(f, g(\theta))$$

Thus QMLE estimates the best-fitting density, where best is measured in terms of the KLIC.

From conventional asymptotic theory, we know

$$\sqrt{n} \left(\hat{\theta}_{QMLE} - \theta_0 \right) \to_d N(0, V)$$

$$
\begin{aligned}
V &= Q^{-1}\Omega Q^{-1} \\
Q &= -E\frac{\partial^2}{\partial\theta\partial\theta'}\log g(y,\theta) \\
\Omega &= E\left(\frac{\partial}{\partial\theta}\log g(y,\theta)\frac{\partial}{\partial\theta}\log g(y,\theta)'\right)
\end{aligned}
$$

If the model is correctly specified ($g(y,\theta_0) = f(y)$), then $Q = \Omega$ (the information matrix equality). Otherwise $Q \neq \Omega$.

15.3 Expected KLIC

The MLE $\hat{\theta} = \hat{\theta}(\mathbf{y})$ is a function of the data vector $\mathbf{y}$.

The fitted model at any $\tilde{\mathbf{y}}$ is $\hat{\mathbf{g}}(\tilde{\mathbf{y}}) = \mathbf{g}(\tilde{\mathbf{y}}, \hat{\theta}(\mathbf{y}))$.

The fitted likelihood is $\mathcal{L}(\hat{\theta}) = -\log \mathbf{g}(\mathbf{y}, \hat{\theta}(\mathbf{y}))$ (the model evaluated at the observed data).

The KLIC of the fitted model is is

$$
\begin{aligned}
KLIC(\mathbf{f}, \hat{\mathbf{g}}) &= C_f - \int \mathbf{f}(\tilde{\mathbf{y}}) \log \mathbf{g}(\tilde{\mathbf{y}}, \hat{\theta}(\mathbf{y}))d\tilde{\mathbf{y}} \\
&= C_f - E_{\tilde{\mathbf{y}}} \log \mathbf{g}(\tilde{\mathbf{y}}, \hat{\theta}(\mathbf{y}))
\end{aligned}
$$

where $\tilde{\mathbf{y}}$ has density $\mathbf{f}$, independent of $\mathbf{y}$.

The expected KLIC is the expectation over the observed values $\mathbf{y}$

$$
\begin{aligned}
E\left(KLIC(\mathbf{f}, \hat{\mathbf{g}})\right) &= C_f - E_{\mathbf{y}} E_{\tilde{\mathbf{y}}} \log \mathbf{g}(\tilde{\mathbf{y}}, \hat{\theta}(\mathbf{y})) \\
&= C_f - E_{\tilde{\mathbf{y}}} E_{\mathbf{y}} \log \mathbf{g}(\mathbf{y}, \hat{\theta}(\tilde{\mathbf{y}}))
\end{aligned}
$$

the second equality by symmetry. In this expression, $\tilde{\mathbf{y}}$ and $\mathbf{y}$ are independent vectors each with density $\mathbf{f}$. Letting $\tilde{\theta} = \hat{\theta}(\tilde{\mathbf{y}})$, the estimator of θ when the data is $\tilde{\mathbf{y}}$, we can write this more compactly as

$$
E_{\cdot y}\left(KLIC(\mathbf{f}, \hat{\mathbf{g}})\right) = C_f - E \log \mathbf{g}(\mathbf{y}, \tilde{\theta})
$$

where $\mathbf{y}$ and $\tilde{\theta}$ are independent.

An alternative interpretation is in terms of predicted likelihood. The expected KLIC is the expected likelihood when the sample $\tilde{\mathbf{y}}$ is used to construct the estimate $\tilde{\theta}$, and an independent sample $\mathbf{y}$ used for evaluation. In linear regression, the quasi-likelihood is Gaussian, and the expected KLIC is the expected squared prediction error.

15.4 Estimating KLIC

We want an estimate of the expected KLIC.

As C_f is constant across models, it is ignored.

We want to estimate

$$
T = -E \log \mathbf{g}(\mathbf{y}, \tilde{\theta})
$$

Make a second-order Taylor expansion of $-\log \mathbf{g}\left(\mathbf{y}, \tilde{\theta}\right)$ about $\hat{\theta}$:

$$\begin{aligned} -\log \mathbf{g}(\mathbf{y}, \tilde{\theta}) &\simeq -\log \mathbf{g}(\mathbf{y}, \hat{\theta}) - \frac{\partial}{\partial \theta} \log \mathbf{g}(\mathbf{y}, \hat{\theta})' \left(\tilde{\theta} - \hat{\theta}\right) \\ &\quad -\frac{1}{2}\left(\tilde{\theta} - \hat{\theta}\right)' \left(\frac{\partial^2}{\partial \theta \partial \theta'} \log \mathbf{g}(\mathbf{y}, \hat{\theta})\right) \left(\tilde{\theta} - \hat{\theta}\right) \end{aligned}$$

The first term on the RHS is $\mathcal{L}(\hat{\theta})$, the second is linear in the FOC, so only the third term remains. Writing

$$\hat{Q} = -n^{-1} \frac{\partial^2}{\partial \theta \partial \theta'} \log \mathbf{g}(\mathbf{y}, \hat{\theta}),$$

$$\tilde{\theta} - \hat{\theta} = \left(\tilde{\theta} - \theta_0\right) - \left(\hat{\theta} - \theta_0\right)$$

and expanding the quadratic, we find

$$-\log \mathbf{g}(\mathbf{y}, \tilde{\theta}) \simeq \mathcal{L}(\hat{\theta}) + \frac{1}{2} n \left(\tilde{\theta} - \theta_0\right)' \hat{Q} \left(\tilde{\theta} - \theta_0\right) + \frac{1}{2} n \left(\hat{\theta} - \theta_0\right)' \hat{Q} \left(\hat{\theta} - \theta_0\right) + n \left(\tilde{\theta} - \theta_0\right) \hat{Q} \left(\hat{\theta} - \theta_0\right).$$

Now

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d Z_1 \sim N(0, V)$$

$$\sqrt{n}\left(\tilde{\theta} - \theta_0\right) \to_d Z_2 \sim N(0, V)$$

which are independent, and $\hat{Q} \to_p Q$. Thus for large n,

$$-\log \mathbf{g}(\mathbf{y}, \tilde{\theta}) \simeq \mathcal{L}(\hat{\theta}) + \frac{1}{2} Z_1' Q Z_1 + \frac{1}{2} Z_2' Q Z_2 + Z_1' Q Z_2.$$

Taking expectations

$$\begin{aligned} T &= -E \log \mathbf{g}(\mathbf{y}, \tilde{\theta}) \\ &\simeq E\mathcal{L}(\hat{\theta}) + E\left(\frac{1}{2} Z_1' Q Z_1 + \frac{1}{2} Z_2' Q Z_2 + Z_1' Q Z_2\right) \\ &= E\mathcal{L}(\hat{\theta}) + \operatorname{tr}(QV) \\ &= E\mathcal{L}(\hat{\theta}) + \operatorname{tr}\left(Q^{-1}\Omega\right) \end{aligned}$$

An (asymptotically) unbiased estimate of T is then

$$\hat{T} = \mathcal{L}(\hat{\theta}) + \operatorname{tr}\widehat{(Q^{-1}\Omega)}$$

where $\operatorname{tr}\widehat{(Q^{-1}\Omega)}$ is an estimate of $\operatorname{tr}\left(Q^{-1}\Omega\right)$.

15.5 AIC

When $g(x, \theta_0) = f(x)$ (the model is correctly specified) then $Q = \Omega$ (the information matrix equality). Hence

$$\operatorname{tr}\left(Q^{-1}\Omega\right) = k = \dim(\theta)$$

so

$$\hat{T} = \mathcal{L}(\hat{\theta}) + k$$

This is the the Akaike Information Criterion (AIC). It is typically written as $2\hat{T}$, e.g.

$$AIC = 2\mathcal{L}(\hat{\theta}) + 2k$$

AIC is an estimate of the expected KLIC, based on the approximation that g includes the correct model.

Picking a model with the smalled AIC is picking the model with the smallest estimated KLIC. In this sense it is picking is the best-fitting model.

15.6 TIC

Takeuchi (1976) proposed a robust AIC, and is known as the Takeuchi Information Criterion (TIC)

$$TIC = 2\mathcal{L}(\hat{\theta}) + 2\operatorname{tr}\left(\hat{Q}^{-1}\hat{\Omega}\right)$$

where

$$\begin{aligned}
\hat{Q} &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}\log g(y_i, \hat{\theta}) \\
\hat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\log g(y_i, \hat{\theta})\frac{\partial}{\partial\theta}\log g(y_i, \hat{\theta})'\right)
\end{aligned}$$

The does not require that g is correctly specified.

15.7 Comments on AIC and TIC

The AIC and TIC are designed for the likelihood (or quasi-likelihood) context. For proper application, the "model" needs to be a conditional density, not just a conditional mean or set of moment conditions. This is a strength and limitation.

The benefit of AIC/TIC is that it selects fitted models whose densities are close to the true density. This is a broad and useful feature.

The relation of the TIC to the AIC is very similar to the relationship between the conventional and "White" covariance matrix estimators for the MLE/QMLE or LS. The TIC does not appear to be widely appreciated nor used.

The AIC is known to be asymptotically optimal in linear regression (we discuss this below), but in the general context I do not know of an optimality result. The desired optimality would be that if a model is selected by minimizing AIC (or TIC) then the fitted KLIC of this model is asymptotically equivalent to the KLIC of the infeasible best-fitting model.

15.8 AIC and TIC in Linear Regression

In linear regression or projection

$$\begin{aligned} y_i &= X_i'\theta + e_i \\ E\left(X_i e_i\right) &= 0 \end{aligned}$$

AIC or TIC cannot be directly applied, as the density of e_i is unspecified. However, the LS estimator is the same as the Gaussian MLE, so it is natural to calculate the AIC or TIC for the Gaussian quasi-MLE.

The Gaussian quasi-likelihood is

$$\log g_i(\theta) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(y_i - X_i'\beta\right)^2$$

where $\theta = (\beta, \sigma^2)$ and $\sigma^2 = Ee_i^2$. The MLE $\hat{\theta} = (\hat{\beta}, \hat{\sigma}^2)$ is LS. The pseudo-true value β_0 is the projection coefficient $\beta = E\left(X_i X_i'\right)^{-1} E\left(X_i y_i\right)$. If β is $k \times 1$ then the number of parameters is $k+1$.

The sample log-likelihood is

$$2\mathcal{L}(\hat{\theta}) = n\log\left(\hat{\sigma}^2\right) + n\log\left(2\pi\right) + n$$

The second/third parts can be ignored. The AIC is

$$AIC = n\log\left(\hat{\sigma}^2\right) + 2\left(k+1\right).$$

Often this is written

$$AIC = n\log\left(\hat{\sigma}^2\right) + 2k$$

as adding/subtracting constants do not matter for model selection, or sometimes

$$AIC = \log\left(\hat{\sigma}^2\right) + 2\frac{k}{n}$$

as scaling doesn't matter.

Also

$$\begin{aligned}\frac{\partial}{\partial\beta}\log g(y_i,\theta) &= \frac{1}{\sigma^2}X_i\left(y_i - X_i'\beta\right)\\ \frac{\partial}{\partial\sigma^2}\log g(y_i,\theta) &= -\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}\left(y_i - X_i'\beta\right)^2,\end{aligned}$$

and

$$\begin{aligned}-\frac{\partial^2}{\partial\beta\partial\beta'}\log g(y_i,\theta) &= \frac{1}{\sigma^2}X_iX_i'\\ -\frac{\partial^2}{\partial\beta\partial\sigma^2}\log g(y_i,\theta) &= \frac{1}{\sigma^4}X_i\left(y_i - X_i'\beta\right)\\ -\frac{\partial^2}{\partial\left(\sigma^2\right)^2}\log g(y_i,\theta) &= -\frac{1}{2\sigma^4}+\frac{1}{\sigma^6}\left(y_i - X_i'\beta\right)^2\end{aligned}$$

Evaluated at the pseudo-true values,

$$\begin{aligned}\frac{\partial}{\partial\beta}\log g(y_i,\theta_0) &= \frac{1}{\sigma^2}X_ie_i\\ \frac{\partial}{\partial\sigma^2}\log g(y_i,\theta_0) &= \frac{1}{2\sigma^4}\left(e_i^2-\sigma^2\right),\end{aligned}$$

and

$$\begin{aligned}-\frac{\partial^2}{\partial\beta\partial\beta'}\log g(y_i,\theta_0) &= \frac{1}{\sigma^2}X_iX_i'\\ -\frac{\partial^2}{\partial\beta\partial\sigma^2}\log g(y_i,\theta_0) &= \frac{1}{\sigma^4}X_ie\\ -\frac{\partial^2}{\partial\left(\sigma^2\right)^2}\log g(y_i,\theta_0) &= \frac{1}{2\sigma^6}\left(2\left(y_i - X_i'\beta\right)^2-\sigma^2\right)\end{aligned}$$

Thus

$$\begin{aligned}Q &= -E\begin{bmatrix}\frac{\partial^2}{\partial\beta\partial\beta'}\log g(y_i,\theta_0) & \frac{\partial^2}{\partial\beta\partial\sigma^2}\log g(y_i,\theta_0)\\ \frac{\partial^2}{\partial\sigma^2\partial\beta'}\log g(y_i,\theta_0) & \frac{\partial^2}{\partial\left(\sigma^2\right)^2}\log g(y_i,\theta_0)\end{bmatrix}\\ &= \sigma^{-2}\begin{bmatrix}E\left(X_iX_i'\right) & 0\\ 0 & \frac{1}{2\sigma^2}\end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}\Omega &= E\begin{bmatrix} \frac{\partial}{\partial\beta}\log g(y_i,\theta_0)\frac{\partial}{\partial\beta}\log g(y_i,\theta_0)' & \frac{\partial}{\partial\beta}\log g(y_i,\theta_0)\frac{\partial}{\partial\sigma^2}\log g(y_i,\theta_0) \\ \frac{\partial}{\partial\sigma^2}\log g(y_i,\theta_0)\frac{\partial}{\partial\beta}\log g(y_i,\theta_0)' & \left(\frac{\partial}{\partial\sigma^2}\log g(y_i,\theta_0)\right)^2 \end{bmatrix} \\ &= \sigma^{-2}\begin{bmatrix} E\left(X_iX_i'\frac{e_i^2}{\sigma^2}\right) & \frac{1}{2\sigma^4}E\left(X_i'e_i^3\right) \\ \frac{1}{2\sigma^4}E\left(X_ie_i^3\right) & \frac{\kappa_4}{4\sigma^2} \end{bmatrix}\end{aligned}$$

where

$$\kappa_4 = var\left(\frac{e_i^2}{\sigma^2}\right) = \frac{E\left(e_i^2-\sigma^2\right)^2}{\sigma^4} = \frac{E\left(e_i^4\right)-\sigma^4}{\sigma^4}$$

We see that $\Omega = Q$ if

$$\begin{aligned} E\left(\frac{e_i^2}{\sigma^2} \mid X_i\right) &= 1 \\ E\left(X_ie_i^3\right) &= 0 \\ \kappa_4 &= 2 \end{aligned}$$

Essentially, this requies that $e_i \sim N(0,\sigma^2)$. Otherwise $\Omega \neq Q$.

Thus the AIC is appropriate in Gaussian regression. It is an "approximation" in non-Gaussian regression, heteroskedastic regression, or projection.

To calculate the TIC, note that since Q is block diagonal you do not need to estimate the off-diagonal component of Ω. Note that

$$\begin{aligned}\operatorname{tr} Q^{-1}\Omega &= \operatorname{tr}\left[E\left(X_iX_i'\right)^{-1}E\left(X_iX_i'\frac{e_i^2}{\sigma^2}\right)\right] + \left(\frac{1}{2\sigma^2}\right)^{-1}\frac{\kappa_4}{4\sigma^2} \\ &= \operatorname{tr}\left[E\left(X_iX_i'\right)^{-1}E\left(X_iX_i'\frac{e_i^2}{\sigma^2}\right)\right] + \frac{\kappa_4}{2}\end{aligned}$$

Let

$$\hat{\kappa}_4 = \frac{1}{n}\sum_{i=1}^n\left(\hat{e}_i^2-\hat{\sigma}^2\right)^2$$

The TIC is then

$$\begin{aligned} TIC &= n\log\left(\hat{\sigma}^2\right) + \operatorname{tr}\left(\hat{Q}^{-1}\hat{\Omega}\right) \\ &= n\log\left(\hat{\sigma}^2\right) + 2\left[\operatorname{tr}\left(\left(\sum_{i=1}^n X_iX_i'\right)^{-1}\left(\sum_{i=1}^n X_iX_i'\frac{\hat{e}_i^2}{\hat{\sigma}^2}\right)\right) + \hat{\kappa}_4\right] \\ &= n\log\left(\hat{\sigma}^2\right) + \frac{2}{\hat{\sigma}^2}\sum_{i=1}^n h_i\hat{e}_i^2 + \hat{\kappa}_4\end{aligned}$$

where $h_i = X_i' \left(\mathbf{X}'\mathbf{X}\right)^{-1} X_i$.

When the errors are close to homoskedastic and Gaussian, then h_i and e_i^2 will be uncorrelated $\hat{\kappa}_4$ will be close to 2, so the penalty will be close to

$$2\sum_{i=1}^{n} h_i + 2 = 2\left(k+1\right)$$

as for AIC. In this case TIC will be close to AIC. In applications, the differences will arise under heteroskedasticity and non-Gaussianity.

The primary use of AIC and TIC is to compare models. As we change models, typically the residuals $\hat{e}_i$ do not change too much, so my guess is that the estimate $\hat{\kappa}$ will not change much. In this event, the TIC correction for estimation of σ^2 will not matter much.

15.9 Asymptotic Equivalence

Let $\tilde{\sigma}^2$ be a preliminary (model-free) estimate of σ^2. The AIC is equivalent to

$$\begin{aligned}
\tilde{\sigma}^2\left(AIC - n\log\tilde{\sigma}^2 + n\right) &= n\sigma^2\left(\log\left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right)+1\right) + 2\tilde{\sigma}^2 k \\
&\simeq n\tilde{\sigma}^2\left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right) + 2\tilde{\sigma}^2 k \\
&= \hat{e}'\hat{e} + 2\tilde{\sigma}^2 k \\
&= C_k
\end{aligned}$$

The approximation is $\log(1+a) \simeq a$ for a small. This is the Mallows criterion. Thus AIC is approximately equal to Mallows, and the approximation is close when k/n is small.

Furtheremore, this expression approximately equalts

$$\hat{e}'\hat{e}\left(1 + \frac{2}{nk}\right) = S_k$$

which is known as Shibata's condition (Annals of Statistics, 1980; Biometrick, 1981).

The TIC (ignoring the correction for estimation of σ^2) is equivalent to

$$\begin{aligned}
\tilde{\sigma}^2\left(TIC - n\log\tilde{\sigma}^2 + n\right) &= n\tilde{\sigma}^2\left(\log\left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right)+1\right) + \frac{2\tilde{\sigma}^2}{\hat{\sigma}^2}\sum_{i=1}^{n} h_i\hat{e}_i^2 \\
&\simeq \hat{e}'\hat{e} + 2\sum_{i=1}^{n} h_i\hat{e}_i^2 \\
&\simeq \sum_{i=1}^{n} \frac{\hat{e}_i^2}{\left(1-h_i\right)^2} \\
&= CV,
\end{aligned}$$

the cross-validation criterion. Thus $TIC \simeq CV$.

They are both asymptotically equivalent to a "Heteroskedastic-Robust Mallows Criterion"

$$C_k^* = \hat{e}'\hat{e} + 2\sum_{i=1}^{n} h_i \hat{e}_i^2$$

which, strangely enough, I have not seen in the literature.

15.10 Mallows Criterion

Ker-Chau Li (1987, Annals of Statistics) provided a important treatment of the optimality of model selection methods for homoskedastic linear regression. Andrews (1991, JoE) extended his results to allow conditional heteroskedasticity.

Take the regression model

$$\begin{aligned} y_i &= g(X_i) + e_i \\ &= g_i + e_i \\ E(e_i \mid X_i) &= 0 \\ E\left(e_i^2 \mid X_i\right) &= 0 \end{aligned}$$

Written as an $n \times 1$ vector

$$y = g + e.$$

Li assumed that the X_i are non-random, but his analysis can be re-interpreted by treating everything as conditional on X_i.

Li considered estimators of the $n \times 1$ vector g which are linear in y and thus take the form

$$\hat{g}(h) = M(h)y$$

where $M(h)$ is $n \times n$, a function of the X matrix, indexed by $h \in H$, and H is a discrete set. For example, a series estimator sets $M(h) = X_h \left(X_h'X_h\right)^{-1} X_h$ where X_h is an $n \times k_h$ set of basis functions of the regressors, and $H = \{1, ..., \bar{h}\}$. The goal is to pick h to minimize the average squared error

$$L(h) = \frac{1}{n}\left(g - \hat{g}(h)\right)'\left(g - \hat{g}(h)\right).$$

The index h is selected by minimizing the Mallows, Generalized CV, or CV criterion. We discuss Mallows in detail, as it is the easiest to analyze. Andrews showed that only CV is optimal under heteroskedasticity.

The Mallows criterion is

$$C(h) = \frac{1}{n}\left(y - \hat{g}(h)\right)'\left(y - \hat{g}(h)\right) + \frac{2\sigma^2}{n}\operatorname{tr} M(h)$$

The first term is the residual variance from model h, the second is the penalty. For series estimators, $\operatorname{tr} M(h) = k_h$. The Mallows selected index $\hat{h}$ minimizes $C(h)$.

Since $y = e + g$, then $y - \hat{g}(h) = e + g - \hat{g}(h)$, so

$$\begin{aligned} C(h) &= \frac{1}{n}\left(y - \hat{g}(h)\right)'\left(y - \hat{g}(h)\right) + \frac{2\sigma^2}{n}\operatorname{tr} M(h) \\ &= \frac{1}{n}e'e + L(h) + 2\frac{1}{n}e'\left(g - \hat{g}(h)\right) + \frac{2\sigma^2}{n}\operatorname{tr} M(h) \end{aligned}$$

And

$$\hat{g}(h) = M(h)y = M(h)g + M(h)e$$

then

$$\begin{aligned} g - \hat{g}(h) &= \left(I - M(h)\right)g - M(h)e \\ &= b(h) - M(h)e \end{aligned}$$

where $b(h) = \left(I - M(h)\right)g$, and $C(h)$ equals

$$\frac{1}{n}e'e + L(h) + 2\frac{1}{n}e'b(h) + \frac{2}{n}\left(\sigma^2 \operatorname{tr} M(h) - e'M(h)e\right)$$

As the first term doesn't involve h, it follows that $\hat{h}$ minimizes

$$C^*(h) = L(h) + 2\frac{1}{n}e'b(h) + \frac{2}{n}\left(\sigma^2 \operatorname{tr} M(h) - e'M(h)e\right)$$

over $h \in H$.

The idea is that empirical criterion $C^*(h)$ equals the desired criterion $L(h)$ plus a stochastically small error.

We calculate that

$$\begin{aligned} L(h) &= \frac{1}{n}\left(g - \hat{g}(h)\right)'\left(g - \hat{g}(h)\right) \\ &= \frac{1}{n}b(h)'b(h) - \frac{2}{n}b(h)'M(h)e + \frac{1}{n}e'M(h)'M(h)e \end{aligned}$$

and

$$\begin{aligned} R(h) &= E\left(L(h) \mid \mathbf{X}\right) \\ &= \frac{1}{n}b(h)'b(h) + E\left(\frac{1}{n}e'M(h)'M(h)e \mid \mathbf{X}\right) \\ &= \frac{1}{n}b(h)'b(h) + \frac{\sigma^2}{n}\operatorname{tr} M(h)'M(h) \end{aligned}$$

The optimality result is:

Theorem 1. Let $\lambda_{\max}(A)$ denote the maximum eigenvalue of A. If for some positive integer m,

$$\begin{aligned} \lim_{n\to\infty} \sup_{h\in H} \lambda_{\max}\left(M(h)\right) &< \infty \\ E\left(e_i^{4m} \mid X_i\right) &\leq \kappa < \infty \\ \sum_{h\in H} \left(nR(h)\right)^{-m} &\to 0 \end{aligned} \tag{1}$$

then

$$\frac{L(\hat{h})}{\inf_{h\in H} L(h)} \to_p 1.$$

15.11 Whittle's Inequalities

To prove Theorem 1, Li (1987) used two key inequalities from Whittle (1960, Theory of Probability and Its Applications).

Theorem. Suppose the observations are independent. Let $\mathbf{b}$ be any $n \times 1$ vector and $\mathbf{A}$ any $n \times n$ matrix, functions of $\mathbf{X}$. If for some $s \geq 2$

$$\max_i E\left(|e_i|^s \mid X_i\right) \leq \kappa_s < \infty$$

then

$$E\left(\left|\mathbf{b}'\mathbf{e}\right|^s \mid \mathbf{X}\right) \leq K_{1s}\left(\mathbf{b}'\mathbf{b}\right)^{s/2} \tag{2}$$

and

$$E\left(\left|\mathbf{e}'\mathbf{A}\mathbf{e} - E\left(\mathbf{e}'\mathbf{A}\mathbf{e} \mid \mathbf{X}\right)\right|^s \mid \mathbf{X}\right) \leq K_{2s}\left(\operatorname{tr}\mathbf{A}'\mathbf{A}\right)^{s/2} \tag{3}$$

where

$$K_{1s} = \frac{2^{s3/2}}{\sqrt{\pi}}\Gamma\left(\frac{s+1}{2}\right)\kappa_s$$

$$K_{2s} = 2^s K_{1s} K_{1,2s}^{1/2} \kappa_{2s}$$

15.12 Proof of Theorem 1

The main idea is similar to that of consistent estimation. Recall that if $S_n(\theta) \to_p S(\theta)$ uniformly in θ, then the minimizer of $S_n(\theta)$ converges to the minimizer of $S(\theta)$. We can write the uniform convergence as

$$\sup_\theta \left|\frac{S_n(\theta)}{S(\theta)} - 1\right| \to_p 0$$

In the present case, we will show (below) that

$$\sup_h \left|\frac{C^*(h) - L(h)}{L(h)}\right| \to_p 0 \tag{4}$$

Let h_0 denote the minimizer of $L(h)$. Then

$$\begin{aligned} 0 &\leq \frac{L(\hat{h}) - L(h_0)}{L(\hat{h})} \\ &= \frac{C^*(\hat{h}) - L(h_0)}{L(\hat{h})} - \frac{C^*(\hat{h}) - L(\hat{h})}{L(\hat{h})} \\ &= \frac{C^*(\hat{h}) - L(h_0)}{L(\hat{h})} + o_p(1) \\ &\leq \frac{C^*(h_0) - L(h_0)}{L(\hat{h})} + o_p(1) \\ &\leq \frac{C^*(h_0) - L(h_0)}{L(h_0)} + o_p(1) \\ &= o_p(1) \end{aligned}$$

This uses (4) twice, and the facts $L(h_0) \leq L(\hat{h})$ and $C^*(\hat{h}) \leq C^*(h_0)$. This shows that

$$\frac{L(h_0)}{L(\hat{h})} \to_p 1$$

which is equivalent to the Theorem.

The key is thus (4). We show below that

$$\sup_h \left| \frac{L(h)}{R(h)} - 1 \right| \to_p 0 \tag{5}$$

which says that $L(h)$ and $R(h)$ are asymptotically equivalent, and thus (4) is equivalent to

$$\sup_h \left| \frac{C^*(h) - L(h)}{R(h)} \right| \to_p 0. \tag{6}$$

From our earlier equation for $C^*(h)$, we have

$$\sup_h \left| \frac{C^*(h) - L(h)}{R(h)} \right| \leq 2 \sup_h \frac{|e'b(h)|}{nR(h)} + 2 \sup_h \frac{\left|\sigma^2 \operatorname{tr} M(h) - e'M(h)e\right|}{nR(h)}. \tag{7}$$

Take the first term on the right-hand-side. By Whittle's first inequality,

$$E\left(\left|e'b(h)\right|^{2m} \mid \mathbf{X}\right) \leq K\left(b(h)'b(h)\right)^m$$

Now recall

$$nR(h) = b(h)'b(h) + \sigma^2 \operatorname{tr} M(h)'M(h) \tag{8}$$

Thus

$$nR(h) \geq b(h)'b(h)$$

Hence

$$E\left(\left|e'b(h)\right|^{2m} \mid \mathbf{X}\right) \leq K\left(b(h)'b(h)\right)^m \leq K\left(nR(h)\right)^m$$

Then, since H is discrete, by applying Markov's inequality and this bound,

$$\begin{aligned} P\left(\sup_h \frac{|e'b\,(h)|}{nR(h)} > \delta \mid \mathbf{X}\right) &\leq \sum_{h\in H} P\left(\frac{|e'b(h)|}{nR(h)} > \delta \mid \mathbf{X}\right) \\ &\leq \sum_{h\in H} \delta^{-2m} \frac{E\left(|e'b(h)|^{2m} \mid \mathbf{X}\right)}{\left(nR(h)\right)^{2m}} \\ &\leq \sum_{h\in H} \delta^{-2m} \frac{K\left(nR(h)\right)^m}{\left(nR(h)\right)^{2m}} \\ &= \frac{K}{\delta^{2m}} \sum_{h\in H} \left(nR(h)\right)^{-m} \\ &\to 0 \end{aligned}$$

by assumption (1). This shows

$$\sup_h \frac{|e'b(h)|}{nR(h)} \to_p 0$$

Now take the second term in (7). By Whittle's second inequality, since

$$E\left(e'M(h)e \mid \mathbf{X}\right) = \sigma^2 \operatorname{tr} M(h),$$

then

$$\begin{aligned} E\left(\left|e'M(h)e - \sigma^2 \operatorname{tr} M(h)\right|^{2m} \mid \mathbf{X}\right) &\leq K\left(\operatorname{tr}\left(M(h)'M(h)\right)\right)^m \\ &\leq \sigma^{-2m} K\left(nR(h)\right)^m \end{aligned}$$

the second inequality since (8) implies

$$\operatorname{tr} M(h)'M(h) \leq \sigma^{-2} nR(h)$$

Applying Markov's inequality

$$\begin{aligned}
P\left(\sup_h \frac{\left|e'M(h)e - \sigma^2 \operatorname{tr} M(h)\right|}{nR(h)} > \delta \mid \mathbf{X}\right) &\leq \sum_{h\in H} P\left(\frac{\left|e'M(h)e - \sigma^2 \operatorname{tr} M(h)\right|}{nR(h)} > \delta \mid \mathbf{X}\right) \\
&\leq \sum_{h\in H} \delta^{-2m} \frac{E\left(\left|e'M(h)e - \sigma^2 \operatorname{tr} M(h)\right|^{2m} \mid \mathbf{X}\right)}{(nR(h))^{2m}} \\
&\leq \sum_{h\in H} \delta^{-2m} \frac{\sigma^{-2m} K (nR(h))^m}{(nR(h))^{2m}} \\
&= K\left(\delta^2\sigma^2\right)^{-m} \sum_{h\in H} (nR(h))^{-m} \\
&\to 0
\end{aligned}$$

For completeness, let us show (5). The demonstration is essentially the same as the above. We calculate

$$\begin{aligned}
L(h) - R(h) &= -\frac{2}{n} b(h)'M(h)e + \frac{1}{n} e'M(h)'M(h)e - \frac{\sigma^2}{n} \operatorname{tr} M(h)'M(h) \\
&= -\frac{2}{n} b(h)'M(h)e + \frac{1}{n}\left(e'M(h)'M(h)e - E\left(e'M(h)'M(h)e \mid \mathbf{X}\right)\right)
\end{aligned}$$

Thus

$$\sup_h \left|\frac{L(h) - R(h)}{R(h)}\right| \leq 2 \sup_h \frac{|e'M(h)'b(h)|}{nR(h)} + 2 \sup_h \frac{|e'M(h)'M(h)e - E\left(e'M(h)'M(h)e \mid \mathbf{X}\right)|}{nR(h)}.$$

By Whittle's first inequality,

$$E\left(\left|e'M(h)'b(h)\right|^{2m} \mid \mathbf{X}\right) \leq K\left(b(h)'M(h)M(h)'b(h)\right)^m$$

Use the matrix inequality

$$\operatorname{tr}(AB) \leq \lambda_{\max}(A) \operatorname{tr}(B)$$

and letting

$$\bar{M} = \lim_{n\to\infty} \sup_{h\in H} \lambda_{\max}\left(M(h)\right) < \infty$$

then

$$\begin{aligned}
b(h)'M(h)M(h)'b(h) &= \operatorname{tr}\left(M(h)M(h)'b(h)b(h)'\right) \\
&\leq \bar{M}^2 \operatorname{tr}\left(b(h)b(h)'\right) \\
&\leq \bar{M}^2 b(h)'b(h) \\
&\leq \bar{M}^2 nR(h)
\end{aligned}$$

Thus

$$\begin{aligned} E\left(\left|e'M(h)'b(h)\right|^{2m} \mid \mathbf{X}\right) &\leq K\left(b(h)'M(h)M(h)'b(h)\right)^m \\ &\leq K\bar{M}^2\left(nR(h)\right)^m \end{aligned}$$

Thus

$$\begin{aligned} P\left(\sup_h \frac{|e'M(h)'b\left(h\right)|}{nR(h)} > \delta \mid \mathbf{X}\right) &\leq \sum_{h\in H} P\left(\frac{|e'M(h)'b(h)|}{nR(h)} > \delta \mid \mathbf{X}\right) \\ &\leq \sum_{h\in H} \delta^{-2m}\frac{E\left(|e'M(h)'b(h)|^{2m} \mid \mathbf{X}\right)}{\left(nR(h)\right)^{2m}} \\ &\leq \sum_{h\in H} \delta^{-2m}\frac{K\bar{M}^2\left(nR(h)\right)^m}{\left(nR(h)\right)^{2m}} \\ &= \frac{K\bar{M}^2}{\delta^{2m}}\sum_{h\in H}\left(nR(h)\right)^{-m} \\ &\to 0 \end{aligned}$$

Similarly,

$$\begin{aligned} E\left(\left|e'M(h)'M(h)e - E\left(e'M(h)'M(h)e \mid \mathbf{X}\right)\right|^{2m} \mid \mathbf{X}\right) &\leq K\left(\mathrm{tr}\left(M(h)'M(h)M(h)'M(h)\right)\right)^m \\ &\leq K\bar{M}^{2m}\left(\mathrm{tr}\left(M(h)'M(h)\right)\right)^m \\ &\leq \sigma^{-2m}K\bar{M}^{2m}\left(nR(h)\right)^m \end{aligned}$$

and thus

$$\begin{aligned} P\left(\sup_h \frac{|e'M(h)'M(h)e - E\left(e'M(h)'M(h)e \mid \mathbf{X}\right)|}{nR(h)} > \delta \mid \mathbf{X}\right) &\leq \sum_{h\in H} P\left(\frac{|e'M(h)'M(h)e - E\left(e'M(h)'M(h)e \mid \mathbf{X}\right.}{nR(h)}\right. \\ &\leq \sum_{h\in H} \delta^{-2m}\frac{E\left(|e'M(h)'M(h)e - E\left(e'M(h)'M(h)e\right.\right.}{\left(nR(h)\right)^{2m}} \\ &\leq \sum_{h\in H} \delta^{-2m}\frac{\sigma^{-2m}K\bar{M}^{2m}\left(nR(h)\right)^m}{\left(nR(h)\right)^{2m}} \\ &= K\left(\frac{\bar{M}^2}{\delta^2\sigma^2}\right)^m\sum_{h\in H}\left(nR(h)\right)^{-m} \\ &\to 0 \end{aligned}$$

We have shown

$$\sup_h\left|\frac{L(h)-R(h)}{R(h)}\right| \to_p 0$$

which is (5).

15.13 Mallows Model Selection

Li's Theorem 1 applies to a variety of linear estimators. Of particular interest is model selection (e.g. series estimation).

Let's verify Li's conditions, which were

$$\begin{aligned} \lim_{n\to\infty} \sup_{h\in H} \lambda_{\max}\left(M(h)\right) &< \infty \\ E\left(e_i^{4m} \mid X_i\right) &\leq \kappa < \infty \\ \sum_{h\in H} \left(nR(h)\right)^{-m} &\to 0 \end{aligned} \tag{9}$$

In linear estimation, $M(h)$ is a projection matrix, so $\lambda_{\max}\left(M(h)\right) = 1$ and the first equation is automatically satisfied.

The key is equation (9).

Suppose that for sample size n, there are N_n models. Let

$$\xi_n = \inf_{h\in H} nR(h)$$

and assume

$$\xi_n \to \infty$$

A crude bound is

$$\sum_{h\in H} \left(nR(h)\right)^{-m} \leq N_n \xi_n^{-m}$$

If $N_n \xi_n^{-m} \to 0$ then (9) holds. Notice that by increasing m, we can allow for larger N_n (more models) but a tighter moment bound.

The condition $\xi_n \to 0$ says that for all finite models h, the there is non-zero approximation error, so that $R(h)$ is non-zero. In contrast, if there is a finite dimensional model h_0 for which $b(h_0) = 0$, then $nR(h_0) = h_0\sigma^2$ does not diverge. In this case, Mallows (and AIC) are asymptotically sub-optimal.

We can improve this condition if we consider the case of selection among models of increasing size. Suppose that model h has k_h regressors, and $k_1 < k_2 < \cdots$ and for some $m \geq 2$,

$$\sum_{h=1}^{\infty} k_h^{-m} < \infty$$

This includes nested model selection, where $k_h = h$ and $m = 2$. Note that

$$nR(h) = b(h)'b(h) + k_h\sigma^2 \geq k_h\sigma^2$$

Now pick $B_n \to \infty$ so that $B_n \xi_n^{-m} \to 0$ (which is possible since $\xi_n \to \infty$.) Then

$$\begin{aligned}\sum_{h=1}^{\infty} (nR(h))^{-m} &= \sum_{h=1}^{B_n} (nR(h))^{-m} + \sum_{h=B_n+1}^{\infty} (nR(h))^{-m} \\ &\leq B_n \xi_n^{-m} + \sigma^{-2} \sum_{h=B_n+1}^{\infty} k_h^{-m} \to 0\end{aligned}$$

as required.

15.14 GMM Model Selection

This is an underdeveloped area. I list a few papers.

Andrews (1999, Econometrica). He considers selecting moment conditions to be used for GMM estimation. Let p be the number of parameters, c represent a list of "selected" moment conditions, $|c|$ denote the cardinality (number) of these moments, and $J_n(c)$ the GMM criterion computed using these c moments. Andrews' proposes criteria of the form

$$IC(c) = J_n(c) - r_n(|c| - p)$$

where $|c| - p$ is the number of overidentifying restrictions and r_n is a sequence. For an AIC-like criterion, he sets $r_n = 2$, for a BIC-like criterion, he sets $r_n = \log n$.

The model selection rule picks the moment conditions c which minimize $J_n(c)$.

Assuming that a subset of the moments are incorrect, Andrews shows that the BIC-like rule asymptotically selects the correct subset.

Andrews and Lu (2001, JoE) extend the above analysis to the case of jointly picking the moments and the parameter vector (that is, imposing zero restrictions on the parameters). They show that the same criterion has similar properties – that it can asymptotically select the "correct" moments and "correct" zero restrictions.

Hong, Preston and Shum (ET, 2003) extend the analysis of the above papers to empirical likelihood. They show that that this criterion has the same interpretation when $J_n(c)$ is replaced by the empirical likelihood.

These papers are an interesting first step, but they do not address the issue of GMM selection when the true model is potentially infinite dimensional and/or misspecified. That is, the analysis is not analogous to that of Li (1987) for the regression model.

In order to properly understand GMM selection, I believe we need to understand the behavior of GMM under misspecification.

Hall and Inoue (2003, Joe) is one of the few contributions on GMM under misspecificaiton. They did not investigate model selection.

Suppose that the model is

$$m(\theta) = Em_i(\theta) = 0$$

where m_i is $\ell \times 1$ and θ is $k \times 1$. Assume $\ell > r$ (overidentification). The model is misspecified if there is no θ such that this moment condition holds. That is, for all θ,

$$m(\theta) \neq 0$$

Suppose we apply GMM. What happens?

The first question is, what is the pseudo-true value? The GMM criterion is

$$J_n(\theta) = n\bar{m}_n(\theta)' W_n \bar{m}_n(\theta)$$

If $W_n \to_p W$, then

$$n^{-1} J_n(\theta) \to_p m(\theta)' W m(\theta).$$

Thus the GMM estimator $\hat{\theta}$ is consistent for the pseudo-true value

$$\theta_0(W) = \operatorname{argmin} m(\theta)' W m(\theta).$$

Interstingly, the pseudo-true value $\theta_0(W)$ is a function of W. This is a fundamental difference from the correctly specified case, where the weight matrix only affects efficiency. In the misspecified case, it affects what is being estimated.

This means that when we apply "iterated GMM", the pseudo-true value changes with each step of the iteration!

Hall and Inoue also derive the distribution of the GMM estimator. They find that the distribution depends not only on the randomness in the moment conditions, but on the randomness in the weight matrix. Specifically, they assume that $n^{1/2}(W_n - W) \to_d Normal$, and find that this affects the asymptotic distributions.

Furthermore, the distribution of test statistics is non-standard (a mixture of chi-squares). So inference on the pseudo-true values is troubling.

This subject deserves more study.

15.15 KLIC for Moment Condition Models Under Misspecification

Suppose that the true density is $f(y)$, and we have an over-identified moment condition model, e.g. for some function $m(y)$, the model is

$$Em(y) = 0$$

However, we want to allow for misspecification, namely that

$$Em(y) \neq 0$$

To explore misspefication, we have to ask: What is a desirable pseudo-true model?

Temporarily ignoring parameter estimation, we can ask: Which density $g(y)$ satisfying this moment condition is closest to $f(y)$ in the sense of minimizing KLIC? We can call this $g_0(y)$ the pseudo-true density.

The solution is nicely explained in Appendix A of Chen, Hong, and Shum (JoE, 2007). Recall

$$KLIC(f,g) = \int f(y) \log\left(\frac{f(y)}{g(y)}\right) dy$$

The problem is

$$\min_g KLIC(f,g)$$

subject to

$$\begin{aligned} \int g(y)dy &= 1 \\ \int m(y)g(y)dy &= 0 \end{aligned}$$

The Lagrangian is

$$\int f(y) \log\left(\frac{f(y)}{g(y)}\right) dy + \mu\left(\int g(y)dy - 1\right) + \lambda' \int m(y)g(y)dy$$

The FOC with respect to $g(y)$ at some y is

$$0 = -\frac{f(y)}{g(y)} + \mu + \lambda' m(y)$$

Multiplying by $g(y)$ and integrating,

$$\begin{aligned} 0 &= -\int f(y)dy + \mu \int g(y)dy + \lambda' \int m(y)g(y)dy \\ &= -1 + \mu \end{aligned}$$

so $\mu = 1$. Solving for $g(y)$ we find

$$g(y) = \frac{f(y)}{1 + \lambda' m(y)},$$

a tilted version of the true density $f(y)$. Inserting this solution we find

$$KLIC(f,g) = \int f(y) \log\left(1 + \lambda' m(y)\right) dy$$

By duality, the optimal Lagrange multiplier λ_0 maximizes this expression

$$\lambda_0 = \underset{\lambda}{\text{argmax}} \int f(y) \log\left(1 + \lambda' m(y)\right) dy.$$

The pseudo-true density is

$$g_0(y) = \frac{f(y)}{1 + \lambda_0' m(y)},$$

with associated minimized KLIC

$$\begin{aligned} KLIC\left(f, g_0\right) &= \int f(y) \log\left(1 + \lambda_0' m(y)\right) dy \\ &= E \log\left(1 + \lambda_0' m(y)\right) \end{aligned}$$

This is the smallest possible KLIC(f,g) for moment condition models.

This solution looks like empirical likelihood. Indeed, EL minimizes the empirical KLIC, and this connection is widely used to motivate EL.

When the moment $m(y, \theta)$ depends on a parameter θ,then the pseudo-true values (θ_0, λ_0) are the joint solution to the problem

$$\min_{\theta} \max_{\lambda} E \log\left(1 + \lambda' m(y, \theta)\right)$$

Theorem (Chen, Hong and Shum, JoE, 2007). If $|m(y, \theta)|$ is bounded, then the EL estimates $(\hat{\theta}, \hat{\lambda})$ are $n^{-1/2}$ consistent for the pseudo-true values (θ_0, λ_0).

This gives a simple interpretation to the definition of KLIC under misspecification.

15.16 Schennach's Impossibility Result

Schennach (Annals of Statistics, 2007) claims a fundamental flaw in the application of KLIC to moment condition models. She shows that the assumption of bounded $|m(y, \theta)|$ is not merely a technical condition, it is binding.

[Notice: In the linear model, $m(y, \theta) = z(y - x'\theta)$ is unbounded if the data has unbounded support. Thus the assumption is highly relevant.]

The key problem is that for any $\lambda \neq 0$, if $m(y, \theta)$ is unbounded, so is $1 + \lambda' m(y, \theta)$. In particular, it can take on negative values. Thus $\log\left(1 + \lambda' m(y, \theta)\right)$ is ill-defined. Thus there is no pseudo-true value of λ. (It must be non-zero, but it cannot be non-zero!) Without a non-zero λ, there is no way to define a pseudo-true θ_0 which satisfies the moment condition.

Technically, Schennach shows that when there is no θ such that $Em(y, \theta) = 0$ and $m(y, \theta)$ is unbounded, then there is no θ_0 such that $\sqrt{n}\left(\hat{\theta} - \theta_0\right) = O_p(1)$.

Her paper leaves open the question: For what is $\hat{\theta}$ consistent? Is there a pseudo-true value? One possibility is that the pseudo-true value θ_n needs to be indexed by sample size. (This idea is used in Hal White's work.)

Never-the-less, Schennach's theorem suggests that empirical likelihood is non-robust to misspecification.

15.17 Exponential Tilting

Instead of

$$KLIC(f,g) = \int f(y) \log\left(\frac{f(y)}{g(y)}\right) dy$$

consider the reverse distance

$$KLIC(g,f) = \int g(y) \log\left(\frac{g(y)}{f(y)}\right) dy.$$

The pseudo-true g which minimizes this criterion is

$$\min_g \int g(y) \log\left(\frac{g(y)}{f(y)}\right) dy$$

subject to

$$\begin{aligned} \int g(y)dy &= 1 \\ \int m(y)g(y)dy &= 0 \end{aligned}$$

The Lagrangian is

$$\int g(y) \log\left(\frac{g(y)}{f(y)}\right) dy - \mu\left(\int g(y)dy - 1\right) - \lambda' \int m(y)g(y)dy$$

with FOC

$$0 = \log\left(\frac{g(y)}{f(y)}\right) + 1 - \mu - \lambda' m(y).$$

Solving

$$g(y) = f(y) \exp(-1+\mu) \exp\left(\lambda' m(y)\right).$$

Imposing $\int g(y)dy = 1$ we find

$$g(y) = \frac{f(y) \exp\left(\lambda' m(y)\right)}{\int f(y) \exp\left(\lambda' m(y)\right) dy}. \tag{10}$$

Hence the name "exponential tilting" or ET

Inserting this into $KLIC(g,f)$ we find

$$\begin{aligned} KLIC(g,f) &= \int g(y)\log\left(\frac{\exp\left(\lambda' m(y)\right)}{\int f(y)\exp\left(\lambda' m(y)\right)dy}\right)dy \\ &= \lambda'\int m(y)g(y)dy - \int g(y)dy\log\left(\int f(y)\exp\left(\lambda' m(y)\right)dy\right) \\ &= -\log\left(\int f(y)\exp\left(\lambda' m(y)\right)dy\right) \qquad (11) \\ &= -\log E\exp\left(\lambda' m(y)\right) \qquad (12) \end{aligned}$$

By duality, the optimal Lagrange multiplier λ_0 maximizes this expression, equivalently

$$\lambda_0 = \underset{\lambda}{\operatorname{argmin}}\, E\exp\left(\lambda' m(y)\right) \qquad (13)$$

The pseudo-true density $g_0(y)$ is (10) with this λ_0, with associated minimized KLIC (11). This is the smallest possible KLIC(g,f) for moment condition models.

Notice: the g_0 which minimize KLIC(g,f) and KLIC(f,g) are different.

In contrast to the EL case, the ET problem (13) does not restrict λ, and there are no "trouble spots". Thus ET is more robust than EL. The pseudo-true λ_0 and g_0 are well defined under misspecification, unlike EL.

When the moment $m(y,\theta)$ depends on a parameter θ,then the pseudo-true values (θ_0,λ_0) are the joint solution to the problem

$$\max_{\theta}\min_{\lambda} E\exp\left(\lambda' m(y,\theta)\right).$$

15.18 Exponential Tilting – Estimation

The ET or exponential tilting estimator solves the problem

$$\min_{\theta,p_1,\ldots,p_n}\sum_{i=1}^{n} p_i\log p_i$$

subject to

$$\begin{aligned} \sum_{i=1}^{n} p_i &= 1 \\ \sum_{i=1}^{n} p_i m\left(y_i,\theta\right) &= 0 \end{aligned}$$

First, we concentrate out the probabilities. For any θ, the Lagrangian is

$$\sum_{i=1}^{n} p_i \log p_i - \mu \left(\sum_{i=1}^{n} p_i - 1 \right) - \lambda' \sum_{i=1}^{n} p_i m\left(y_i, \theta\right)$$

with FOC

$$0 = \log \hat{p}_i - 1 - \mu - \lambda' m(y_i, \theta).$$

Solving for $\hat{p}_i$ and imposing the summability,

$$\hat{p}_i(\lambda) = \frac{\exp\left(\lambda' m(y_i, \theta)\right)}{\sum_{i=1}^{n} \exp\left(\lambda' m(y_i, \theta)\right)}$$

When $\lambda = 0$ then $\hat{p}_i = n^{-1}$, same as EL. The concentrated "entropy" criterion is then

$$\begin{aligned} \sum_{i=1}^{n} \hat{p}_i(\lambda) \log \hat{p}_i(\lambda) &= \sum_{i=1}^{n} \hat{p}_i(\lambda) \left[\lambda' m(y_i, \theta) - \log \left(\sum_{i=1}^{n} \exp\left(\lambda' m(y_i, \theta)\right) \right) \right] \\ &= -\log \left(\sum_{i=1}^{n} \exp\left(\lambda' m(y_i, \theta)\right) \right) \end{aligned}$$

By duality, the Lagrange multiplier maximizes this criterion, or equivalently

$$\hat{\lambda}(\theta) = \underset{\lambda}{\operatorname{argmin}} \sum_{i=1}^{n} \exp\left(\lambda' m(y_i, \theta)\right)$$

The ET estimator $\hat{\theta}$ maximizes this concentrated function, e.g.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \exp\left(\hat{\lambda}(\theta)' m(y_i, \theta)\right)$$

The ET probabilities are $\hat{p}_i = \hat{p}_i(\hat{\lambda})$

15.19 Schennach's Estimator

Schennach (2007) observed that while the ET probabilities have desirable properties, the EL estimator for θ has better bias properties. She suggested a hybrid estimator which achieves the best of both worlds, called exponentially tilted empirical likelihood (ETEL).

This is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, ELET(\theta)$$

$$
\begin{aligned}
ETEL(\theta) &= \sum_{i=1}^{n} \log\left(\hat{p}_i(\theta)\right) \\
&= \hat{\lambda}(\theta)' \sum_{i=1}^{n} m(y_i,\theta) - \log\left(\sum_{i=1}^{n} \exp\left(\hat{\lambda}(\theta)' m(y_i,\theta)\right)\right)
\end{aligned}
$$

$$
\hat{p}_i(\theta) = \frac{\exp\left(\hat{\lambda}(\theta)' m(y_i,\theta)\right)}{\sum_{i=1}^{n} \exp\left(\hat{\lambda}(\theta)' m(y_i,\theta)\right)}
$$

$$
\hat{\lambda}(\theta) = \underset{\lambda}{\operatorname{argmin}} \sum_{i=1}^{n} \exp\left(\lambda' m(y_i,\theta)\right)
$$

She claims the following advantages for the ETEL estimator $\hat{\theta}$

- Under correct specification, $\hat{\theta}$ is asymptotically second-order equivalent to EL
- Under misspecification, the pseudo-true values λ_0, θ_0 are generically well defined, and minimize a KLIC analog
- $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d N\left(0, \Gamma^{-1}\Omega\Gamma^{-1\prime}\right)$ where $\Gamma = E\dfrac{\partial}{\partial\theta'} m\left(y,\theta\right)$ and $\Omega = E\; m\left(y,\theta\right) m\left(y,\theta\right)'$.