1 Comparative Studies

I would like to review widely used standard methods for the model selection. Throughout, $\mathcal{M}_i$ stands for a model.

1.0.1 AIC, BIC

Bayesian Information Criterion (BIC) derives from Laplace approximation of predictive model distribution. Suppose $M_i \sim P(\theta)$. Then the predictive likelihood for the training dataset $x = \{x_1, ..., x_n\}$ is given by

$$\begin{aligned} p(x|M) &= \int p(x|\theta) p_{M_i}(\theta) d\theta \\ &= \int \exp\left(\log\left(p(x_n|\theta) p_{M_i}(\theta)\right)\right) d\theta \end{aligned} \tag{1.1}$$

The laplace approxmation claims that, for "regular" f,

$$\lim_{n\to\infty} \frac{\int_a^b \exp(nf(x))dx}{\sqrt{\frac{2\pi}{n|\partial_x^2 f(x_0)|}} \exp(f(x_0)))} = 1 \tag{1.2}$$

where $x_0 = \arg\sup_{x\in[a,b]} f(x)$. The principle of the proof is shown in the derivation of FIC to be provided below, which is just the taylor expansion followed by Gaussian integral. The BIC yields that

$$\log P(y|\hat{\theta}_i) - \log P(y|M_i) \in O\left(\dim(\theta)\log \mathrm{n}\right) \tag{1.3}$$

And this allows us to choose model based on posterior of the parameter evaluated at the model's optimal. AIC, on the other hand, is an application of central limit theorem. In reality, we always wish to evaluate θ by computing

$$\int p(y) \log p(y|\theta) dy \tag{1.4}$$

Indeed, however, the exact computation is impossible unless $p(y)$ is known exactly. We taylor expand $\log p(y|\theta)$, and conduct central limit theorem about maximum likelihood estimator. The AIC is gives us that

$$\frac{1}{n}\sum_{k=1}^{n} \log p(y_k|\hat{\theta}) - \int p(y) \log p(y|\hat{\theta}) dy \in O\left(\frac{d}{n}\right) \tag{1.5}$$

1.0.2 WBIC

1.0.3 ARD

1.1 FIC, FAB

1.1.1 VB vs FAB

We begin with the model

$$Y = (X \cdot Z)\beta + \epsilon$$

where Z is the mask, that is, $Z_k \sim \text{Bernnouli}(\mathrm{\pi_k})$, $\epsilon \sim \mathcal{N}(0, \lambda^{-1} I)$ and X is given. Put $\theta = (\beta, \lambda) \sim p(\cdot|\theta_0)$. In the approach of variational inference, one will try to approximate the joint distribution of all hidden variables by independent random variables:

$$p(Z, \theta) \cong q(\theta) \prod_i q(Z_{nk}|\mu_{nk})^{Z_{nk}}$$

where μ_{nk} now depends on n as well. Still explained other way, we do the minimization of

$$\begin{aligned}
\log p(Y|X, \pi) &= E_{q(Z,\theta)} \left[\log p(Y|X, \pi)\right] \\
&= E_{q(Z,\theta)} \left[\log \frac{p(Y, Z, \theta|Y, X, \pi)}{p(Z, \theta|X, \pi)}\right] \\
&= E_{q(Z,\theta)} \left[\log p(Y, Z, \theta|X, \pi)\right] - E_{q(Z,\theta)}[\log q(Z, \theta)] \\
&\qquad + \mathrm{KL[q(Z, \theta)\|p(Z, \theta|Y, X, \pi)]} \\
&\geq E_{q(Z,\theta)} \left[\log p(Y, Z, \theta|X, \pi)\right] - E_{q(Z,\theta)}[\log q(Z, \theta)]
\end{aligned} \tag{1.6}$$

In standard textbooks like PRML, they explain the sequential minimization of the the objective function above about q that puts

$$\log \tilde{p}(\theta) = E_{q(Z)}[\log p(Y, Z, \theta|X, \pi)]$$

$$\log \tilde{p}(Z) = E_{q(\theta)}[\log p(Y, Z, \theta|X, \pi)]$$

in order to express (1.6) the objective function as Kullback Leibler divergences between $\tilde{p}$ and q . FIC questions this paradigm of separating θ from Z. In particular, with exactly same argument as above we can get

$$\log p(Y|X, \pi) \geq E_{q(Z)} \left[\log p(Y, Z|X, \pi)\right] - E_{q(Z)}[\log q(Z)] \tag{1.7}$$

where

$$p(Y, Z|X, \pi) = \int_{\Omega_\theta} p(Y, Z, \theta|X, \pi) d\theta \tag{1.8}$$

By applying 2nd order Taylor expansion we obtain

$$\begin{aligned}\log p(Y,Z,\theta|X,\pi) &= \log p(Y,Z|\theta,X,\pi) + \log p(\theta) &(1.9)\\ &\cong \log p(Y,Z,|\hat{\theta},X,\pi) - \frac{1}{2}N[F_\theta^N(\hat{\theta}),(\theta-\hat{\theta})] + \log p(\theta) &(1.10)\end{aligned}$$

where we use the notation $[A,x]=x^TAx$ and

$$F_\theta^N(\hat{\theta}) = -\frac{1}{N}\partial_\theta^2 \log p(Y,Z,|\hat{\theta},X,\pi)\Big|_{\hat{\theta}}$$

This way, we can include the correlation(2nd moment information) between θ and Z. Appealing to law of large numbers, we get

$$F_\theta^N(\hat{\theta}) \to \int p(Z,|\theta,X,\pi)\partial_\theta^2(\log p(Y,Z|\theta,X,\pi))dZd\theta\Big|_{\hat{\theta}} := F_\theta(\hat{\theta})$$

Therefore, for large enough N,

$$\begin{aligned}\int_{\Omega_\theta} p(Y,Z,\theta,|X,\pi)d\theta &= p(Y,Z|X,\hat{\theta},\pi)\int_{\Omega_\theta}\exp\left(-\frac{N}{2}[F_\theta^N(\hat{\theta}),(\theta-\hat{\theta})]\right)p(\theta)d\theta &(1.11)\\ &\cong p(Y,Z|X,\hat{\theta},\pi)\int_{\Omega_\theta}\exp\left(-\frac{N}{2}[F_\theta(\hat{\theta}),(\theta-\hat{\theta})]\right)p(\theta)d\theta &(1.12)\\ &= p(Y,Z|X,\hat{\theta},\pi)\sqrt{\frac{(2\pi)^{\dim(\theta)}}{N^{\dim(\theta)}|F_\theta(\hat{\theta})|}}E_\Theta[p(\Theta)] &(1.13)\end{aligned}$$

where Θ is Gaussian with variance $NF_\theta(\hat{\theta})$. Putting this back into (1.7) we get

$$\begin{aligned}\log p(Y|X,\pi) \geq E_{q(Z)}\left[\log p(Y,Z|X,\hat{\theta},\pi)\right] - \frac{1}{2}E_{q(Z)}\left[\log\left(\det(\mathrm{F}_\theta(\hat{\theta}))\right)\right] &&(1.14)\\ -\frac{\dim(\theta)}{2}\log(N) - E_{q(Z)}[\log q(Z)] + \log\left(E_\Theta[p(\Theta)]\right) &&(1.15)\end{aligned}$$

The right hand side is defined as FIC. Note that $F_\beta(XZ)^T(XZ)$. The Hadamard's inequality claims that, if H is a matrix, then

$$\det(\mathrm{N}) \leq \prod_{\mathrm{i}=1}^{\mathrm{n}}\|\mathrm{H_i}\| \qquad (1.16)$$

where H_i are the column vectors of H. Therefore **if we ignore** F_λ, we get

$$\log\left(\det(\mathrm{F}_\theta(\hat{\theta}))\right) \leq \log\left(\sum x_{nk}^2 \sum z_{nk}^2\right) \leq \log\sum z_{nk} + \mathrm{const} \qquad (1.17)$$

because $z_{nk} \in \{0, 1\}$. And finally, the convexity of log dictates that

$$\log\left(\sum z_{nk}\right) \leq \log N\pi_k + \frac{1}{N\pi_k}\left(\left(\sum z_{nk}\right) - N\pi_k\right) \tag{1.18}$$

and we reach the lower bound

$$\log p(Y|X,\pi) \geq E_{q(Z)}\left[\log p(Y, Z|\hat{\theta}, X, \pi)\right] - \frac{1}{2}\sum_k \left(\log N\pi_k + \frac{1}{N\pi_k}\left(\left(\sum_n \mu_{nk}\right) - N\pi_k\right)\right)$$
$$- \frac{\dim(\theta)}{2}\log(N) - E_{q(Z)}[\log q(Z)] + \log\left(E_\Theta[p(\Theta)]\right) \tag{1.19}$$

Taking the advantage of the fact that all terms grows with N except the last term, the original Fujimaki's work ignores the last term in the lower bound for large value of N. This might not be a smart move for the case of dim(θ) >> N. The objective function that we will aim to optimize is therefore

$$\mathcal{G}(\mu, \hat{\theta}, \pi) := \log p(Y|X,\pi) \geq E_{q(Z)}\left[\log p(Y|Z, \hat{\theta}, X, \pi)\right] + E_{q(Z)}\left[\log p(Z|\pi)\right] \tag{1.20}$$

$$- \frac{1}{2}\sum_k \left(\log N\pi_k + \frac{1}{N\pi_k}\left(\left(\sum_n \mu_{nk}\right) - N\pi_k\right)\right) \tag{1.21}$$

$$- \frac{\dim(\theta)}{2}\log(N) - E_{q(Z)}[\log q(Z)] \tag{1.22}$$

$$\tag{1.23}$$

2 Optimization of (1.19)

We will optimize the lower bound by sequentially optimizing the RHS with respect to q, or equivalently μ and $\hat{\theta}, \pi$. The former is referred to as E step and the latter as M step. When I took the derivative of the above with respect to μ_{nk} I was able to obtain the same answer as Bayesian Masking paper. At this point, I will just buy their computation.

3 Appendix

3.1