Now consider \( K > 2 \):

\[
p(c_k | \mathbf{x}) = \frac{e^{a_k}}{\sum_{j=1}^{K} e^{a_j}}
\]

\[
d_k = \mathbf{w}_k^T \cdot \mathbf{x}
\]

The goal is to determine \( \{ \mathbf{w}_k \} \) directly.

Use \( \mathbf{t}_n = \{ t_{n,1}, \ldots, t_{n,K} \} \)

\[
\begin{align*}
T & = \left( \begin{array}{c}
\vdots \\
T_{nk} = t_{nk}
\end{array} \right) \\
& = \left( \begin{array}{c}
\vdots \\
y_n(z_n)
\end{array} \right)
\end{align*}
\]

Then

\[
L = p(T | \mathbf{w}_1, \ldots, \mathbf{w}_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(c_k | \mathbf{x}_n) t_{nk} = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}
\]

\[
E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\log L = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log y_{nk}
\]

Now, consider

\[
\frac{\partial E}{\partial \mathbf{w}_j} = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{1}{y_{nk}} \frac{\partial y_{nk}}{\partial \mathbf{w}_j}
\]

\[
\text{vector derivative}
\]

Note that \( \frac{\partial y_{nk}}{\partial w_j} = 0 \), \( j \neq j' \).
\[
\frac{\partial y_k}{\partial a_{ij}} = y_k \delta_{kj} - \frac{e_{dk}}{(\sum_{j'} j' e_{aj'})^2} e_{ai}
= y_k \delta_{kj} - y_k y_j.
\]

\[
- \sum_{n,k} \frac{t_{nk}}{y_{nk}} \left[ y_{nk} \delta_{kj} - y_{nk} y_{nj} \right] \bar{y}_n =
\]

\[
= \sum_{n,k} \left[ t_{nk} y_{nj} - t_{nk} \delta_{kj} \right] \bar{y}_n = \sum_n \left[ y_{nj} - t_{nj} \right] \bar{y}_n.
\]

Moreover,

\[
H = \nabla^2 E = \sum_n \left[ \frac{\partial^2 y_{nj}}{\partial \bar{y}_k \partial \bar{y}_j} \right] \bar{y}_n
= \sum_n \left[ y_{nj} \delta_{kj} - y_{nk} y_{nj} \right] \bar{y}_n \bar{y}_n^T
\]

Can show that \( \bar{u}^T H \bar{u} > 0 \) \( \Rightarrow \) minimum unique positive definite.

Thus can do NR algorithm again.
Probit regression

If desired, \( G(a) \) can be replaced by

\[
P(a) = \int_{-\infty}^{a} d\phi \quad N(\theta_0, 1)
\]

cumulative gaussian, or

probit function

Indeed, consider \( G(a) = \frac{1}{1+e^{-a}} \) vs. \( \phi(\lambda a) \).

\[
\frac{dG(a)}{da} \bigg|_{a=0} = G(0) (1-G(0)) = \frac{1}{4}.
\]

But \[
\frac{d\phi(\lambda a)}{da} \bigg|_{a=0} = \frac{\lambda}{\sqrt{2\pi}} e^{-\left(\lambda a\right)^2} \bigg|_{a=0} = \frac{\lambda}{\sqrt{2\pi}}.
\]

Match derivatives at \( a=0 \):

\[
\frac{2}{\sqrt{2\pi}} = \frac{1}{4} \quad \Rightarrow \quad \lambda^2 = \frac{\pi}{8}.
\]

So, \( G(a) \approx \phi\left(\frac{\sqrt{\pi}}{2\sqrt{2}} a\right) \).

Can repeat the ML analysis for the logistic regression with probit regression \( \Rightarrow \) similar results in practice.
Consider

\[ p(z) = \frac{f(z)}{\mathcal{Z}} \]

\[ \int dz \, p(z) = 1 \implies \mathcal{Z} = \int dz \, f(z). \]

\[ f(z) = f(z_0) e^{-\frac{A}{2} (z-z_0)^2} \]

\[ A \equiv -\frac{\partial^2}{\partial z^2} \log f(z) \bigg|_{z_0} = \frac{\partial f(z)}{\partial z} \bigg|_{z_0} = 0 \]

\[ \log f(z) \approx \log f(z_0) - \frac{A}{2} (z-z_0)^2 \]

Then

\[ p(z) \implies q(z) = \sqrt{\frac{A}{2\pi}} e^{-\frac{A}{2} (z-z_0)^2} = N(z, A^{-1}) \]

Likewise,

\[ f(z) \approx f(z_0) e^{-\frac{A}{2} (z-z_0)^T A^{-1} (z-z_0)} \]

\[ p(\tilde{z}) \implies q(\tilde{z}) = \frac{1}{(2\pi)^{D/2}} e^{-\frac{1}{2} (\tilde{z}-\tilde{z}_0)^T A^{-1} (\tilde{z}-\tilde{z}_0)} \]

\[ = N(\tilde{z} | \tilde{z}_0, A^{-1}) \]

Note that

\[ z = \int d\tilde{z} \, f(\tilde{z}) = f(z_0) \left( \frac{1}{2\pi} \right)^{D/2} \frac{1}{|A|^{1/2}} \]

\[ \approx \frac{f(z_0)}{(2\pi)^{D/2}} e^{-\frac{1}{2} (\tilde{z}-\tilde{z}_0)^T A^{-1} (\tilde{z}-\tilde{z}_0)} \]

\[ \approx \frac{f(z_0)}{(2\pi)^{D/2}} e^{-\frac{1}{2} (\tilde{z}-\tilde{z}_0)^T A^{-1} (\tilde{z}-\tilde{z}_0)} \]
Model comparison

Consider a set of models \{M_i\} with prms \{\Theta_i\}. Define likelihood

\[ p(D | \Theta_i, M_i), \]

then

\[ p(D | M_i) = \int d\Theta_i p(D | \Theta_i, M_i) p(\Theta_i | M_i) \]

model evidence

Under saddle-point approximation,

\[ f(\tilde{\Theta}) \Rightarrow p(D | \tilde{\Theta}_i, M_i) p(\tilde{\Theta}_i | M_i) : \]

\[ p(D | M_i) \approx p(D | \tilde{\Theta}_i, \text{MAP}, M_i) p(\tilde{\Theta}_i, \text{MAP} | M_i) \]

\[ \times \frac{(2\pi)^{M/2}}{|A_i|^{1/2}}, \text{ where } \begin{cases} \text{in model } M_i \\ \text{strictly speaking, } M \rightarrow M_i \end{cases} \]

M is the # model prms and

\[ A_i = -\nabla_{\tilde{\Theta}} \nabla_{\tilde{\Theta}} \left[ \log p(D | \tilde{\Theta}_i, M_i) p(\tilde{\Theta}_i, \text{MAP}) \right] \bigg| \tilde{\Theta}_i, \text{MAP} = \]

\[ -\nabla_{\tilde{\Theta}} \nabla_{\tilde{\Theta}} \left[ \log p(\tilde{\Theta}_i, D, M_i) \right] \bigg| \tilde{\Theta}_i, \text{MAP}, \posterior \]

Finally,

\[ \log Z \]

\[ \left[ \log p(D | M_i) \approx \log p(D | \tilde{\Theta}_i, \text{MAP}, M_i) + \right. \]

\[ + \log p(\tilde{\Theta}_i, \text{MAP} | M_i) + \frac{M}{2} \log(2\pi) - \frac{1}{2} \log |A_i| \]

"penalty" for model complexity
Moreover, assume that priors are given by
\[ \text{p}(\tilde{\Theta}_i | M_i) = \mathcal{N}(\tilde{\Theta}_i | \tilde{\Theta}_i, \mathbf{L}_i) \]

Then
\[ A_i = - \tilde{\Theta}_i \mathbf{L}_i^{-1} \mathbf{L}_i^{-1} \frac{\partial}{\partial \hat{\Theta}_i} \left[ \log \text{p}(D | \hat{\Theta}_i, M_i) \right] \bigg|_{\hat{\Theta}_i, \text{MAP}} \]

\[ - \tilde{\Theta}_i \mathbf{L}_i^{-1} \frac{\partial}{\partial \hat{\Theta}_i} \left[ \log \text{p}(\tilde{\Theta}_i | M_i) \right] \bigg|_{\tilde{\Theta}_i, \text{MAP}} = \mathbf{H} + \mathbf{L}_i^{-1} \mathbf{L}_i \]

If \( \mathbf{L}_i \) is small or we have lots of data (\( N \) is large),
\[ - \log \text{p}(D | \hat{\Theta}_i, M_i) \]
has \( \sum_{n=1}^{N} \) and \( \frac{1}{N} \) as \( N \to \infty \)

We have:
\[ \log \text{p}(D | M_i) = \log \text{p}(D | \hat{\Theta}_i, \text{MAP}, M_i) - \frac{1}{2} \text{log}(2\pi) + \frac{M}{2} \text{log}(d) + \frac{M}{2} \text{log}(2\pi) \]

\[ - \frac{1}{2} \hat{\Theta}_i, \text{MAP} \hat{\Theta}_i, \text{MAP} - \frac{1}{2} \log |\mathbf{H}| \log \text{in the large-N limit} \]

Same argument as above (\( \mathbf{L}_i \) small and/or \( N \) large)

Finally,
\[ H = \sum_{n=1}^{N} H_n = N \langle H \rangle \Rightarrow \log |\mathbf{H}| = \log |\mathbf{N}^{\frac{1}{N}} \mathbf{H}| = \frac{1}{N} \sum_{n=1}^{N} H_n \]

\[ = M \log N + \log |\mathbf{H}^N| \]

We obtain:
\[ [\log \text{p}(D | M_i) = \log \text{p}(D | \hat{\Theta}_i, \text{MAP}, M_i) - \frac{M}{2} \log N.] \]

Bayesian information criterion (BIC)

Complexity penalty