4. Further formal analysis

We have already indicated that the logistic model allows for the introduction of new parameters, which may represent, for example, interactions between pairs of existing query terms or the addition of new terms to the query. A superficial analysis suggests that any additional parameters should be useful in helping to distinguish relevant fron non-relevant documents, or at worst neutral. However, some of our early experiments were conducted on term interactions, with entirely negative results.

A deeper analysis indicates, at least in qualitative terms, why this might be. There is only a limited amount of data from which the parameters are to be estimated, and introducing new parameters in effect stretches this data further. Thus all the parameters are less precisely estimated, and this drop in precision more than outweighs (or may do so) the potential improvement due to including the new parameters. This property was indicated previously in van Rijsbergen's (1979) phrase "the curse of dimensionality".

Is it possible to quantify this property, so as to derive a rule or rules which would define when a parameter should or should not be added ? This question is addressed in the present section. We have not in fact succeeded in developing an immediately applicable rule; however, we feel that the analysis gives some insight into the problem, and may also provide a formal basis suitable for further work.

4.1 Definitions

Suppose that we take repeated samples of n documents and the probability of a document in cell j being relevant is constant. If $k_j(t)$ is the number of relevant documents in cell j after t samples then $k_j(t)/t \longrightarrow n_j p_j$ as $t \longrightarrow \infty$ $a.s.$ with the obvious notation

$$\frac{1}{t} f\,(\underline{k}(t), t\underline{n}\,;\,\rho) \longrightarrow \sum_{j=1}^{d} n_j p_j \rho_j - n_j \log\,(1+e^{\rho_j})$$

The value of $\underline{\rho}$ which maximises the limit above will be denoted $\underline{\pi}_{\mathcal{M}}$ $\underline{\pi} - \underline{\pi}_{\mathcal{M}}$ can be thought of as a sort of asymptotic bias. It is where we lose out because our model is wrong. If $\pi \in \mathcal{M}$ then $\underline{\pi}_{\mathcal{M}} = \underline{\pi}$.

As $\mathcal{M}$ is made larger the bias will decrease but the expected value of $\left\| \underline{\pi}_{\mathcal{M}} - \hat{\underline{\pi}}_{\mathcal{M}} \right\|$ will increase.

Suppose we add another vector to $\mathcal{M}$, increasing its dimension by one. We would like to try and get estimates of the way the "bias" $\underline{\pi} - \underline{\pi}_{\mathcal{M}}$ and the "variance" $E\left(\left\| \underline{\pi}_{\mathcal{M}} - \hat{\underline{\pi}}_{\mathcal{M}} \right\|^2 \right)$ vary.

4.2 The "variance"

Let $\psi(x)$ be any real valued function and suppose $\psi(x)$ has a maximum at $x = 0$. Expanding ψ as a power series we have say

$$\psi(x) = a_0 + a_2 x^2 + a_3 x^3 + \ldots\ldots$$

$$\psi'(x) = 2a_2 x + 3a_3 x^2 + \ldots\ldots$$

$$\psi''(x) = 2a_2 + 6a_3 x + \ldots\ldots$$

and so $\psi'(x)/\psi''(x) = x + 0(x^2)$ as $x \longrightarrow 0$.

Now suppose ψ has its maximum at $\hat{x}$. Then if $\left| x - \hat{x} \right|$ is small $\psi'(x)/\psi''(x)$ is a reasonable estimate of $x - \hat{x}$.

When $\underline{x}$ is a vector variable $\psi'(\underline{x})$ becomes the vector of first order partial derivatives and $\psi''(\underline{x})$ the matrix of second order derivatives but the argument still works. Hence we might be able to get some idea of the behaviour of $\underline{\pi}_m - \hat{\underline{\pi}}_m$ by looking at the first and second derivatives of f.

$$\frac{\partial f}{\partial \rho_i} = k_i - n_i \frac{e^{\rho_i}}{1 + e^{\rho_i}} = k_i - n_i r_i \quad \text{say;}$$

with the obvious notation

$$\frac{\partial f}{\partial \underline{\rho}} = \underline{k} - \underline{N}\,\underline{r} .$$

Also

$$\frac{\partial^2 f}{\partial \rho_i \rho_j} = \begin{cases} 0 & i = j \\ n_i - \dfrac{e^{\rho_i}}{(1+e^{\rho_i})^2} = n_i r_i (1-r_i) & i = j \end{cases}$$

which we will abreviate as

$$\frac{\partial^2 f}{\partial \underline{\rho}^2} = \underline{N}\,\underline{\phi}\,(\underline{r}) \text{ where } \underline{\phi}\,(\underline{r}) \text{ is the diagonal matrix whose i'th diagonal entry is } r_i(1 - r_i).$$

Now let $\underline{A}$ be a d x m matrix whose columns are a basis for $\mathcal{M}$ (m being the dimension of $\mathcal{M}$). Writing $\underline{\rho} = \underline{A}\,\underline{\sigma}$ we have (with some abuse of notation)

$$\frac{\partial}{\partial \sigma} f(\underline{A}\,\underline{\sigma}) = \underline{A}^T \frac{\partial f}{\partial \underline{\rho}}(\underline{\rho}) = \underline{A}^T (\underline{k} - \underline{N}\,\underline{r})$$

$$\frac{\partial^2}{\partial \sigma^2} f(\underline{A}\,\underline{\sigma}) = \underline{A}^T \frac{\partial^2 f}{\partial \underline{\rho}^2}(\underline{\rho})\,\underline{A} = \underline{A}^T \underline{N}\,\underline{\phi}\,(\underline{r})\,\underline{A}$$

As both $\underline{\pi}_m$ and $\hat{\underline{\pi}}_m$ are in $\mathcal{M}$ we can find $\underline{\sigma}_1$ and $\underline{\sigma}_2$ such that $\underline{\pi}_m = \underline{A}\,\underline{\sigma}_1$, $\hat{\underline{\pi}}_m = \underline{A}\,\underline{\sigma}_2$ and $\underline{\sigma}_2$ will be the $\underline{\sigma}$ which maximises $f(\underline{A}\,\underline{\sigma})$. Hence we can approximate $\underline{\pi}_m - \hat{\underline{\pi}}_m$ by

$$A\left(\frac{\partial^2}{\partial \underline{\sigma}_1^2} f(\underline{A}\,\underline{\sigma})\right)^{-1} \frac{\partial}{\partial \underline{\sigma}_1} f(\underline{A}\,\underline{\sigma})$$

$$= \underline{A}\,(\underline{A}^T \underline{N}\,\underline{\phi}\,(\underline{p}_m)\,\underline{A})^{-1}\,\underline{A}^T\,(\underline{k} - \underline{N}\,\underline{p}_m)$$

First we note that $\underline{A}^T\,\underline{N}\,\underline{\phi}\,\underline{A}$ will be non-singular if and only if $\underline{N}\,\underline{A}$ has rank m but this is implied by our assumptions that $\underline{N}\,\underline{\mu} = \underline{0}$ and $\underline{\mu} \in \mathcal{M} \Rightarrow \underline{\mu} = \underline{0}$.

Now, if instead of the usual Euclidian distance, we use the norm $\|\underline{\rho}\| = \underline{\rho}^T\,\underline{N}\,\underline{\phi}\,(\underline{p}_m)\,\underline{\rho}$ we find that

$$\|\underline{\pi}_m - \hat{\underline{\pi}}_m\| = (\underline{\pi}_m - \hat{\underline{\pi}}_m)^T\,\underline{N}\,\underline{\phi}\,(\underline{\pi}_m - \hat{\underline{\pi}}_m)$$

$$\sim (\underline{k} - \underline{N}\underline{p}_m)^T \underline{A}\,(\underline{A}^T\underline{N}\,\underline{\phi}\,\underline{A})^{-1}\,\underline{A}^T(\underline{k} - \underline{N}\underline{p}_m) \qquad (1)$$

Because of the way $\underline{p}_m$ is defined we have that $\underline{A}^T(\underline{N}\,\underline{p} - \underline{N}\,\underline{p}_m) = \underline{0}$ and so we can replace our estimate (1) by

$$(\underline{k} - \underline{N}\underline{p})^T\,\underline{A}(\underline{A}^T\underline{N}\,\underline{\phi}\,\underline{A})^{-1}\,\underline{A}^T\,(\underline{k} - \underline{N}\underline{p}) \qquad (2)$$

Now for each j, $k_j - n_j p_j$ is a random variable with mean 0. If for each j we define

$$X_j = \begin{cases} (k_j - n_j p_j)/(n_j p_j (1 - p_j))^{\frac{1}{2}} & n_j = 0 \\ \text{a } N(0,1) \text{ random variable independent of the other } X_i & n_j = 0 \end{cases}$$

then for all j $X_j(n_j p_j (1-p_j))^{\frac{1}{2}} = k_j - p_j n_j$

and we can replace our estimate by

$$\underline{X}^T(\underline{N}\ \underline{\phi}\ (\underline{p}))^{\frac{1}{2}}\ \underline{A}(\underline{A}^T\underline{N}\ \underline{\phi}\ (\underline{p}_{\mathfrak{M}})\underline{A})^{-1}\ \underline{A}^T(\underline{N}\ \underline{\phi}\ (\underline{p}))^{\frac{1}{2}}\underline{X} \qquad (3)$$

Now

$$E(X_i X_j) = \begin{cases} 1 & i = j \\ 0 & i = j \end{cases}$$

and so the expectation of (3) given N will be the trace of the inner matrix. Reorganising it a bit we get

$$E((\underline{\pi}_{\mathfrak{M}} - \underline{\hat{\pi}}_{\mathfrak{M}})^T N\phi\ (p_{\mathfrak{M}})(\underline{\pi}_{\mathfrak{M}} - \underline{\hat{\pi}}_{\mathfrak{M}}) \mid \underline{N})\quad Tr(\underline{A}^T\underline{N}\ \underline{\phi}(\underline{p}\quad)\underline{A})^{-1}$$
$$\sim Tr(\underline{A}^T\underline{N}\ \underline{\phi}\ (\underline{p}_{\mathfrak{M}})\underline{A})^{-1} \qquad (\underline{A}^T\underline{N}\ \underline{\phi}\ (\underline{p})\underline{A}) \qquad (4)$$

The first point about this is that if $\underline{p}_{\mathfrak{M}} = \underline{p}$ then the above estimate equals m (the dimension of $\mathfrak{M}$) and each new vector we add increases it by 1. We will come back to the more general case later.

There are two main questions about the above estimate. "How good is it ?" and "Can we do better ?". It is possible that it is not too bad. The difference between f'/f" and $\underline{\pi} - \underline{\hat{\pi}}$ might average out when we take the expectation. Clearly more work could be done.

4.3 The "bias"

Let $\mathfrak{M}$ and $\underline{A}$ be defined as before and suppose we increase to include the vector $\underline{\alpha}$. We write

$$\mathfrak{N} = \mathfrak{M} \oplus \langle\alpha\rangle \quad \text{and} \quad \underline{B} = (\underline{A}\ \underline{\alpha})$$

We assume that both maxima exist.

i.e.

$$f(\underline{\rho}) = \sum_{j=1}^{d} n_j \left\{ p_j \rho_j - \log(1 + e^{\rho_j}) \right\}$$

has a finite maximum in $\mathcal{N}$ at $\underline{\pi}_n$ and in $\mathcal{M}$ at $\underline{\pi}_m$. $\underline{\pi}_m$ will be the unique vector satisfying.

$$\underline{\pi}_m \in \mathcal{M} \quad \text{and} \quad \underline{A}^T f'(\underline{\pi}_m) = \underline{0}$$

but $f'(\underline{\pi}_m) = \underline{N}(\underline{p} - \underline{p}_m)$ and so $\underline{\pi}_m$ is that unique vector satisying

(i) $\underline{\pi}_m \in \mathcal{M}$

(ii) $\underline{A}^T \underline{N}(\underline{p} - \underline{p}_m) = \underline{0}$

similarly for $\underline{\pi}_n$ we have

(iii) $\underline{\pi}_n \in \mathcal{N}$

(iv) $\underline{B}^T \underline{N}(\underline{p} - \underline{p}_n) = \underline{0}$

Note that (i) $\Longrightarrow$ (iii) and (iv) $\Longrightarrow$ (ii).

Now let us suppose we have an arc $\underline{\gamma}(t)$ in $\mathbb{R}^d$ with the following properties (as usual $\underline{g}(t) = \dfrac{e^{\underline{\gamma}(t)}}{1 + e^{\underline{\gamma}(t)}}$):

(1) $\underline{\gamma}(0) = \underline{\pi}_m$ and $\underline{\gamma}(1) = \underline{\pi}_n$

(2) $\underline{A}^T \underline{N}(\underline{p} - \underline{g}(t)) = 0, \quad 0 \leq t \leq 1$

(3) $\underline{\gamma}(t) \in \mathcal{N}, \quad 0 \leq t \leq 1$

(4) $\underline{\alpha}^T \underline{N}(\underline{p} - \underline{g}(t))$ is monotonously decreasing for $0 \leq t \leq 1$

By (3) we can write $\underline{\gamma}(t) = \underline{B}\,\underline{\zeta}(t)$ $\quad \underline{\zeta}(t) \in \mathbb{R}^{m+1}$

and by (2)

$$\underline{B}^T \underline{N}\,(\underline{p} - \underline{g}(t)) = \underline{\alpha}^T \underline{N}(\underline{p} - \underline{g}(t))\underline{\eta} \text{ where } \underline{\eta} = \begin{pmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix} \begin{matrix} m+1 \\ \text{dim.} \end{matrix}$$

differentiating both sides w.r.t. t gives

$$\frac{d}{dt}\left[\underline{\alpha}^T \underline{N}(\underline{p} - \underline{g}(t))\right]\underline{\eta} = \underline{B}^T \underline{N} \frac{d}{dt} \frac{exp\ \underline{B}\,\underline{\zeta}\,(t)}{1 + exp\ \underline{B}\,\underline{\zeta}\,(t)}$$

$$= \underline{B}^T \underline{N} \frac{e^{\underline{B}\,\underline{\zeta}\,(t)}}{(1 + e^{\underline{B}\,\underline{\zeta}\,(t)})^2} \underline{B}\,\underline{\zeta}'(t)$$

$$= \underline{B}^T \underline{N}\,\underline{\phi}\,(\underline{g}(t))\,\underline{B}\,\underline{\zeta}'\,(t)$$

but $\underline{\gamma}'\ (t) = \underline{B}\,\underline{\zeta}'(t)$ and so

$$\underline{\gamma}'(t) = \underline{B}(\underline{B}^T \underline{N}\,\underline{\phi}\,(\underline{g})\underline{B})^{-1}\,\underline{\eta}\,\frac{d}{dt}\left[\underline{\alpha}^T \underline{N}(\underline{p} - \underline{g}(t))\right]$$

Suppose we can consider that the variation in $\underline{\phi}\,(g)$ is fairly small compared with its size. Then we can write.

$$\underline{\pi}_m - \underline{\pi}_n = \underline{\gamma}\,(0) - \underline{\gamma}\,(1)\quad \underline{B}(\underline{B}^T \underline{N}\,\underline{\phi}(\underline{p}_m)\underline{B})^{-1}\,\underline{\eta}\,\underline{\alpha}^T \underline{N}(\underline{p} - \underline{p}_m)$$

Now if we choose α s.t. $\underline{A}^T \underline{N}\,\underline{\phi}\,(\underline{p}_m)\,\underline{\alpha} = \underline{0}$

$$\underline{B}(\underline{B}^T \underline{N}\,\underline{\phi}\,\underline{B})^{-1}\,\underline{\eta} = \underline{B}\begin{pmatrix} (\underline{A}^T \underline{N}\,\underline{\phi}\,\underline{A})^{-1} & 0 \\ 0 & (\underline{\alpha}^T \underline{N}\,\underline{\phi}\,\underline{\alpha})^{-1} \end{pmatrix}\underline{\eta} = \underline{B}(\underline{\alpha}^T \underline{N}\,\underline{\phi}\,\underline{\alpha})^{-1}\,\underline{\eta}$$

$$= (\underline{\alpha}^T \underline{N}\,\underline{\phi}\,\underline{\alpha})^{-1}\alpha$$

giving

$$\underline{\pi}_m - \underline{\pi}_n \simeq \frac{\underline{\alpha}^T \underline{N}(\underline{p} - \underline{p}_m)}{\underline{\alpha}^T \underline{N}\, \underline{\phi}\, \underline{\alpha}} \underline{\alpha}$$

Hence

$$\underline{\pi}_m - \underline{\pi}_n)^T \underline{N}\, \underline{\phi}\, (\underline{\pi}_m - \underline{\pi}_n) \simeq \frac{\underline{\alpha}^T \underline{N}(\underline{p} - \underline{p}_m)^2}{\underline{\alpha}^T \underline{N}\, \underline{\phi}\, \underline{\alpha}}$$

If we go back to forumla (4) in the "variance" section and impose the conditions $\underline{\alpha}^T \underline{N}\, \underline{\phi}\, \underline{A} = \underline{0}$ and $\underline{\phi}(\underline{p}_m) \simeq \underline{\phi}(\underline{p}_n)$ we find that the change in the variance going from m to n is about $(\underline{\alpha}^T \underline{N}\, \underline{\Phi}\, (\underline{p})\underline{\alpha})/(\underline{\alpha}^T \underline{N}\, \underline{\phi}\, (\underline{p}_m)\underline{\alpha})$

Comparing these we get the criterion that we should include $\underline{\alpha}$ if

$$(\underline{\alpha}^T \underline{N}(\underline{p} - \underline{p}_m))^2 > (\underline{\alpha}^T \underline{N}\, \underline{\phi}\, (\underline{p})\underline{\alpha})$$

4.4 An Example

It is interesting to look at the case when we have only two terms. Then $d = 4$, the 4 classes being

class 0 : all the documents with neither term

class 1 : the documents with term 1 but not term 2

class 2 : the documents with term 2 but not term 1

class 3 : the documents with both terms

We want to see what happens when we add the interaction to the independence model. The independence model m has dimension 3. A suitable basis for m would be the columns of $\underline{A}$ where

$$\underline{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

If we assume that there are some documents in each class (i.e. $\underline{N}$ is non-singular) then a suitable candidate for $\underline{\alpha}$ is $\underline{\alpha} = (\underline{N}\underline{\phi}_m)^{-1}\ \underline{\eta}$ where

$$\eta = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$$

The criterion for the inclusion of $\underline{\alpha}$ then becomes

$$(\underline{\eta}^T\ \underline{\phi}_m^{-1}\ (\underline{p} - \underline{p}_m))^2 > (\underline{\eta}^T(\underline{N}\ \underline{\phi}_m)^{-2}\underline{N}\ \underline{\phi}\ \underline{\eta})$$

If we assume $\underline{\phi} - \underline{\phi}_m$ is small compared to ϕ we can simplify this to

$$(\underline{\eta}^T\underline{\phi}\ (\underline{p})(\underline{p} - \underline{p}_m))^2 > \sum_{j=1}^{4} (p_j(1 - p_j))n_j^{-1}$$

The way the left-hand side varies with $\underline{N}$ needs more investigation but it should be sandwiched between positive upper and lower bounds. The RHS on the other hand tends to 0 only if all the n_j tend to infinity.

The tentative conclusion we might draw from this is that there is no point in including an interaction unless there are a number of documents in each of the four classes.