Final solutions
ECE 271
Electrical and Computer Engineering
University of California San Diego

Nuno Vasconcelos Fall 2013

1.a) Since the π_j are probabilities, we need to solve

$$\Pi^* = \arg\max_{\pi_1,\ldots,\pi_N} \log P_{C_1,\ldots,C_N}(c_1,\ldots,c_N)$$

subject to the constraint

$$\sum_j \pi_j = 1.$$

We use a Lagrangian formulation, where instead of maximizing only the likelihood, we maximize the Lagrangian

$$\begin{aligned} L &= \log P_{C_1,\ldots,C_N}(c_1,\ldots,c_N) + \lambda(\sum_j \pi_j - 1) \\ &= \sum_j c_j \log \pi_j + \log(n!) - \sum_k \log(c_k!) + \lambda(\sum_j \pi_j - 1) \end{aligned}$$

Taking the gradient with respect to both the π_j and λ and setting to zero, we obtain

$$\begin{aligned} \frac{\partial L}{\partial \pi_j} &= \frac{c_j}{\pi_j} + \lambda = 0 \\ \frac{\partial L}{\partial \lambda} &= \sum_j \pi_j - 1 = 0 \end{aligned}$$

Multiplying the first equation by π_j on both sides, summing over j, and using the second equation

$$\lambda = -\sum_j c_j = -n$$

from which it follows (still from the first equation) that

$$\pi_j^* = \frac{c_j}{n}.$$

Computing the Hessian

$$\frac{\partial^2 L}{\partial \pi_j^2} = -\frac{c_j}{\pi_j^2} \qquad \frac{\partial^2 L}{\partial \pi_j \pi_k} = 0$$

it follows that $\nabla^2_{\mathbf{\Pi}} L = -diag(c_j/\pi_j^2)$. This is clearly negative definite since, for a diagonal matrix, the eigenvalues are just the entries in the main diagonal. Hence we have a maximum.

b) The MAP estimate is

$$\begin{aligned} \mathbf{\Pi}_{MAP} &= \arg\max_{\pi_1,\ldots,\pi_N} P_{\mathbf{C}|\mathbf{\Pi}}(c_1,\ldots,c_N|\pi_1,\ldots,\pi_N) P_{\mathbf{\Pi}}(\pi_1,\ldots,\pi_N) \\ &= \arg\max_{\pi_1,\ldots,\pi_N} Z \prod_{j=1}^N \pi_j^{c_j} \prod_{j=1}^N \pi_j^{u_j - 1} \\ &= \arg\max_{\pi_1,\ldots,\pi_N} \prod_{j=1}^N \pi_j^{c_j + u_j - 1} \end{aligned}$$

where Z is a normalizing constant that does not depend on the π_i. Notice that this maximization is exactly the one that we did in **a)** to compute the ML estimate. In fact the MAP estimate is identical to the ML estimate for the case where we have $n + \sum_j u_j - N$ observations from X and the number of the times that the observed value of X is j is $c_j + u_j - 1$. Hence, we can simply recycle the solution of **a)** and conclude that

$$\pi_j{}^* = \frac{c_j + u_j - 1}{n + \sum_j u_j - N} = \frac{c_j + u_j - 1}{\sum_j (c_j + u_j - 1)}$$

c) The MAP solution is equivalent to the ML solution of an equivalent problem with a larger number of trials. In particular we have $\sum_j c_j$ observations with relative counts c_j (real observations) and $\sum_j (u_j - 1)$ observations with relative counts $u_j - 1$ (virtual observations). The prior allows us to force the estimate to go in a direction that we believe to be more plausible. For example, if we set $\sum_j (u_j - 1) >> \sum_j c_j$ and set one of the u_js very close to $\sum_j u_j$ we are basically forcing π_j to go to 1. Conversely, if we set $\sum_j (u_j - 1) >> \sum_j c_j$ and set all the u_js to the same value, we force the π_js to be identical, i.e. we have a uniform distribution. To understand the effect of the prior for large n, we start by noting that

$$\lim_{n\to\infty} \frac{u_j - 1}{n + \sum_j u_j - N} = 0.$$

Hence

$$\lim_{n\to\infty} \pi_{j\,MAP} = \lim_{n\to\infty} \frac{c_j}{n + \sum_j (u_j - 1)}$$

and

$$\lim_{n\to\infty} \frac{\pi_{j\,ML}}{\pi_{j\,MAP}} = \lim_{n\to\infty} \frac{n + \sum_j u_j - N}{n} = 1.$$

That is, as the number of observations increases, the MAP estimate converges to the ML estimate and therefore becomes independent of the prior itself.

2.a.i) As usual, we maximize the log of the cost. This is the log-likelihood of $\mathcal{D}$ under the predictive distribution, i.e.

$$\mathcal{L} = -\sum_i \frac{(x_i - \mu_n)^2}{2(\sigma^2 + \sigma_n^2)} - \frac{n}{2}\log(2\pi(\sigma^2 + \sigma_n^2)).$$

Setting the derivative wrt μ_0 to zero

$$\begin{aligned}\frac{d\mathcal{L}}{d\mu_0} &= \frac{d\mathcal{L}}{d\mu_n}\frac{d\mu_n}{d\mu_0} \\ &= \sum_i \frac{2(x_i - \mu_n)}{2(\sigma^2 + \sigma_n^2)}(1 - \alpha_n) = 0\end{aligned}$$

we obtain

$$\mu_n^* = \frac{1}{n}\sum_i x_i = \mu_{ML}$$

from which it follows that

$$\mu_{ML} = \alpha_n \mu_{ML} + (1 - \alpha_n)\mu_0^*$$

and $\mu_0^* = \mu_{ML}$.

2.a.ii) Plugging this value in the log-likelihood we have

$$\begin{aligned}\mathcal{L}^*(\sigma_0^2) &= -\frac{n}{2(\sigma^2 + \sigma_n^2)}\frac{1}{n}\sum_i (x_i - \mu_{ML})^2 - \frac{n}{2}\log(2\pi(\sigma^2 + \sigma_n^2)) \\ &\propto -\frac{1}{(\sigma^2 + \sigma_n^2)}\nu - \log(\sigma^2 + \sigma_n^2)\end{aligned}$$

where $\nu = \frac{1}{n}\sum_i (x_i - \mu_{ML})^2$ is the sample variance of $\mathcal{D}$ and we dropped all constants. Setting the derivative wrt σ_0^2 to zero we obtain

$$\begin{aligned}\frac{d\mathcal{L}}{d\sigma_0^2} &= \frac{d\mathcal{L}}{d\sigma_n^2}\frac{d\sigma_n^2}{d\sigma_0^2} \\ &= \left(\frac{1}{(\sigma^2 + \sigma_n^2)^2}\nu - \frac{1}{\sigma^2 + \sigma_n^2}\right)\frac{\sigma^2}{n}\frac{n(\sigma^2 + n\sigma_0^2) - n\sigma_0^2 n}{(\sigma^2 + n\sigma_0^2)^2} \\ &= \frac{\sigma^2}{(\sigma^2 + \sigma_n^2)}\left(\frac{\nu}{(\sigma^2 + \sigma_n^2)} - 1\right)\frac{\sigma^2}{(\sigma^2 + n\sigma_0^2)^2} = 0.\end{aligned}$$

For finite n this holds if $\sigma_0^2 = \infty$ or

$$\nu = \sigma^2 + \sigma_n^2 = \sigma^2\left(1 + \frac{\sigma_0^2}{\sigma^2 + n\sigma_0^2}\right) \Leftrightarrow \frac{\sigma_0^2}{\sigma^2 + n\sigma_0^2} = \frac{\nu}{\sigma^2} - 1 \Leftrightarrow \sigma_0^2 = \frac{\sigma^2\left(\frac{\nu}{\sigma^2} - 1\right)}{1 - n\left(\frac{\nu}{\sigma^2} - 1\right)}.$$

Hence

$$(\sigma_0^*)^2 = \frac{\sigma^2\left(\frac{\nu}{\sigma^2} - 1\right)}{1 - n\left(\frac{\nu}{\sigma^2} - 1\right)}.$$

2.b.i) The complete data log-likelihood is

$$\log P_{x,\mu;\mu_0,\sigma_0^2}(\mathcal{D}, \mu; \mu_0, \sigma_0^2) = \log P_{x|\mu}(\mathcal{D}|\mu) + \log P_{\mu;\mu_0,\sigma_0^2}(\mu; \mu_0, \sigma_0^2)$$

$$
\begin{aligned}
&= -\sum_i \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2) - \frac{(\mu-\mu_0)^2}{2\sigma_0^2} - \frac{1}{2}\log(2\pi\sigma_0^2) \\
&\propto -\frac{1}{2\sigma^2}\left(-2\mu\sum_i x_i + n\mu^2\right) - \frac{1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2) - \frac{1}{2}\log(\sigma_0^2) \\
&\propto -\frac{n}{\sigma^2}\left(-2\mu\mu_{ML} + \mu^2\right) - \frac{1}{\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2) - \log(\sigma_0^2) \\
&= -\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 + 2\left(\frac{n\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\mu - \frac{\mu_0^2}{\sigma_0^2} - \log(\sigma_0^2)
\end{aligned}
$$

where we have dropped constants that do not depend on the prior parameters.

2.b.ii) The Q function is

$$
\begin{aligned}
Q[(\mu_0,\sigma_0^2)|(\mu_0,\sigma_0^2)^{(i)}] &= E_{\mu|x;(\mu_0,\sigma_0^2)^{(i)}}[\log P_{x,\mu;\mu_0,\sigma_0^2}(\mathcal{D},\mu;\mu_0,\sigma_0^2)|\mathcal{D}] \\
&= -\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right) E_{\mu|x;(\mu_0,\sigma_0^2)^{(i)}}[\mu^2|\mathcal{D}] + 2\left(\frac{n\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) E_{\mu|x;(\mu_0,\sigma_0^2)^{(i)}}[\mu|\mathcal{D}] - \frac{\mu_0^2}{\sigma_0^2} - \log(\sigma_0^2) \\
&= -\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right) < \mu^2 > +2\left(\frac{n\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) < \mu > -\frac{\mu_0^2}{\sigma_0^2} - \log(\sigma_0^2) \\
&\propto -\frac{1}{\sigma_0^2}(< \mu^2 > -2\mu_0 < \mu > +\mu_0^2) - \log(\sigma_0^2),
\end{aligned}
$$

with

$$
\begin{aligned}
< \mu^2 > &= E_{\mu|x;(\mu_0,\sigma_0^2)^{(i)}}[\mu^2|\mathcal{D}] \qquad (1) \\
< \mu > &= E_{\mu|x;(\mu_0,\sigma_0^2)^{(i)}}[\mu|\mathcal{D}]. \qquad (2)
\end{aligned}
$$

Note that these are just the first and second order moments of the posterior for μ under a prior of parameters $(\mu_0,\sigma_0^2)^{(i)}$. We know, from the table given with the problem, that

$$
P_{\mu|x;(\mu_0,\sigma_0^2)^{(i)}}(\mu|\mathcal{D}) = G(\mu,\mu_n^{(i)},(\sigma_n^2)^{(i)}).
$$

It follows that

$$
\begin{aligned}
< \mu^2 > &= (\sigma_n^2)^{(i)} + (\mu_n^{(i)})^2 \\
< \mu > &= \mu_n^{(i)}
\end{aligned}
$$

and

$$
\begin{aligned}
< \mu > &= \alpha_n^{(i)}\mu_{ML} + (1-\alpha_n^{(i)})\mu_0^{(i)} \\
< \mu^2 > &= \frac{\sigma^2}{n}\alpha_n^{(i)} + < \mu >^2 \\
\alpha_n^{(i)} &= \frac{n(\sigma_0^2)^{(i)}}{\sigma^2 + n(\sigma_0^2)^{(i)}}
\end{aligned}
$$

These are the computations of the E-step.

2.b.iii) To compute the parameter updates we set the derivatives of the Q function to zero. This leads to

$$
\frac{dQ}{d\mu_0} = \frac{2}{\sigma_0^2}(< \mu > -\mu_0) = 0
$$

$$
\begin{aligned}
\mu_0^{(i+1)} &= <\mu> \\
\frac{dQ}{d\sigma_0^2} &= \frac{1}{(\sigma_0^2)^2}(<\mu^2> -2\mu_0 <\mu> +\mu_0^2) - \frac{1}{\sigma_0^2} \\
&= \frac{1}{(\sigma_0^2)^2}(<\mu^2> - <\mu>^2) - \frac{1}{\sigma_0^2} = 0 \\
(\sigma_0^2)^{(i+1)} &= <\mu^2> - <\mu>^2 .
\end{aligned}
$$

c) From EM equations

$$
\begin{aligned}
\mu_0^{(i+1)} &= \mu_n^{(i)} = \alpha_n^{(i)}\mu_{ML} + (1-\alpha_n^{(i)})\mu_0^{(i)} \\
(\sigma_0^2)^{(i+1)} &= (\sigma_n^2)^{(i)} = \frac{\sigma^2}{n}\alpha_n^{(i)}
\end{aligned}
$$

At convergence we have

$$
\begin{aligned}
\mu_0^* &= \alpha_n^*\mu_{ML} + (1-\alpha_n^*)\mu_0^* \\
(\sigma_0^2)^* &= \frac{\sigma^2}{n}\alpha_n * \\
&= \frac{\sigma^2(\sigma_0^2)^*}{\sigma^2 + n(\sigma_0^2)^*}
\end{aligned}
$$

from which

$$
\begin{aligned}
\mu_0^* &= \mu_{ML} \\
\sigma^2 + n(\sigma_0^2)^* &= \sigma^2 \\
(\sigma_0^2)^* &= 0
\end{aligned}
$$

i.e. the prior converges to a delta function centered at μ_{ML}. Hence

$$\alpha_n^* = 0$$

and the predictive distribution is a Gaussian of mean μ_{ML} and variance σ^2, i.e. the ML model. For predictive-ML, we have

$$
\begin{aligned}
\mu_n^* &= \mu_{ML} \\
\alpha_n^* &= \frac{\frac{n\sigma^2\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)}}{\sigma^2 + \frac{n\sigma^2\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)}} = \frac{\frac{n\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)}}{1 + \frac{n\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)}} = \frac{n\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)} \\
(\sigma_n^2)^* &= \frac{\sigma^2}{n}\alpha_n^* = \frac{\sigma^2\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)}
\end{aligned}
$$

and the predictive distribution is a Gaussian of mean μ_{ML} and variance

$$\sigma^2\left[1 + \frac{\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)}\right].$$

For a large sample ν converges to σ^2 and this is well approximated by σ^2. Hence the two empirical Bayesian procedures produce the same predictive distribution, which is the ML model, for large samples.

This is not surprising because Bayes and ML tend to be equivalent for large samples. To analyze the small sample case, we recall that

$$\begin{aligned}\frac{d\mathcal{L}}{d\sigma_0^2} &= \frac{\sigma^2}{(\sigma^2+\sigma_n^2)}\left(\frac{\nu}{(\sigma^2+\sigma_n^2)}-1\right)\frac{\sigma^2}{(\sigma^2+n\sigma_0^2)^2} \\ &\propto \frac{\nu}{(\sigma^2+\sigma_n^2)}-1\end{aligned}$$

The optimal solution occurs for $\sigma_n^2 = 0$ if and only if $\nu \leq \sigma^2$. (Note that if $\nu < \sigma^2$ the gradient is negative and the optimal solution is $\sigma_n^2 = 0$).

3. a) Denoting by S_k^l the set $\{S_k, \ldots, S_l$, with $l \geq k$,

$$\begin{aligned} P_{S_i|S_0}(s_i|s_0) &= \int P_{S_i, S_1^{i-1}|S_0}(s_i, s_1^{i-1}|s_0) ds_{i-1} \ldots ds_1 \\ &= \int P_{S_i|S_0^{i-1}}(s_i|s_0^{i-1}) P_{S_1^{i-1}|S_0}(s_1^{i-1}|s_0) ds_{i-1} \ldots ds_1 \end{aligned}$$

Using the Markovian property

$$\begin{aligned} P_{S_i|S_0}(s_i|s_0) &= \int P_{S_i|S_{i-1}}(s_i|s_{i-1}) P_{S_1^{i-1}|S_0}(s_1^{i-1}|s_0) ds_{i-1} \ldots ds_1 \\ &= \int P_{S_i|S_{i-1}}(s_i|s_{i-1}) P_{S_{i-1}|S_0^{i-2}}(s_{i-1}|s_0^{i-2}) P_{S_1^{i-2}|S_0}(s_1^{i-2}|s_0) ds_{i-1} \ldots ds_1 \\ &= \int P_{S_i|S_{i-1}}(s_i|s_{i-1}) P_{S_{i-1}|S_{i-2}}(s_{i-1}|s_{i-2}) P_{S_1^{i-2}|S_0}(s_1^{i-2}|s_0) ds_{i-1} \ldots ds_1 \end{aligned}$$

and, applying the same procedure iteratively

$$\begin{aligned} P_{S_i|S_0}(s_i|s_0) &= \int P_{S_i|S_{i-1}}(s_i|s_{i-1}) P_{S_{i-1}|S_{i-2}}(s_{i-1}|s_{i-2}) \ldots P_{S_2|S_1}(s_2|s_1) P_{S_1|S_0}(s_1|s_0) ds_{i-1} \ldots ds_1 \\ &= \int P_{S_i|S_{i-1}}(s_i|s_{i-1}) P_{S_{i-1}|S_{i-2}}(s_{i-1}|s_{i-2}) \ldots \left[\int P_{S_2|S_1}(s_2|s_1) P_{S_1|S_0}(s_1|s_0) ds_1 \right] ds_{i-1} \ldots ds_2 \\ &= \int P_{S_i|S_{i-1}}(s_i|s_{i-1}) \left\{ \int P_{S_{i-1}|S_{i-2}}(s_{i-1}|s_{i-2}) \ldots \left[\int P_{S_2|S_1}(s_2|s_1) P_{S_1|S_0}(s_1|s_0) ds_1 \right] ds_{i-2} \right\} ds_{i-1}. \end{aligned}$$

Using the trick that we have talked about in class

$$\begin{aligned} \int P_{S_2|S_1}(s_2|s_1) P_{S_1|S_0}(s_1|s_0) ds_1 &= \int G(s_2, s_1, \sigma^2) G(s_1, s_0, \sigma^2) ds_1 \\ &= G(s_2, 0, \sigma^2) \star G(s_2, s_0, \sigma^2) \\ &= G(s_2, s_0, 2\sigma^2), \end{aligned}$$

where $\star$ means convolution. Note that each integral is of this form, and is replaced by a Gaussian whose variance increases by σ^2. Hence

$$P_{S_i|S_0}(s_i|s_0) = G(s_i, s_0, i \times \sigma^2).$$

b) This is a standard problem of Bayesian estimation. Note that we have a prior

$$P_{S_0}(s_0) = G(s_0, \mu, \nu^2),$$

and a likelihood, given by the result of **a)**, and the independence assumption

$$P_{S_i|S_0}(\mathcal{D}_i|s_0) = \prod_k G(x_k^i, s_0, i \times \sigma^2).$$

This is the problem of estimating the posterior for the mean of Gaussian, when the prior is a Gaussian, which we studied in class. We have seen that

$$P_{S_0|S_i}(s_0|\mathcal{D}_i) = G(s_0, s_i^{(MAP)}, \hat{\sigma}_i^2)$$

with

$$\begin{aligned} s_i^{(MAP)} &= \frac{n\nu^2}{i\times\sigma^2+n\nu^2}s_i^{(ML)} + \frac{i\times\sigma^2}{i\times\sigma^2+n\nu^2}\mu \\ s_i^{(ML)} &= \frac{1}{n}\sum_k x_k^i \\ \frac{1}{\hat{\sigma}_i^2} &= \frac{1}{\nu^2}+\frac{n}{i\times\sigma^2} \end{aligned}$$

The MAP estimate is s_{MAP}. For small i it is dominated by $s_i^{(ML)}$, the mean of the observations. This means that the estimate is dominated by the received signal, which makes sense since the number of hops is small and not much noise has been added. For large i it is dominated by the prior mean μ. This, once again, makes sense. The signal has gone through many hops and the observations are too noisy to be trusted. In this case, we should stick with the prior.

c) We start by considering the process by which the measurements are collected. Using the variable K to denote the link that was selected, there are two steps

1. technician picks link k with probability $P_K(k)=\frac{1}{L}$.
2. technician measures x_i from that link. If s_0 is known this means that x_i is sampled from $P_{S_k|S_0}(x_i|s_0)$. Otherwise it is sampled from $P_{S_k}(x_i)$.

Overall, the data is sampled from a mixture model. Using the variable X to denote the observation, we have, for case 1

$$P_{X|S_0}(x|s_0) = \sum_k P_{X|K,S_0}(x|k,s_0)P_K(k) = \sum_k P_{S_k|S_0}(x|s_0)P_K(k) = \frac{1}{L}\sum_k G(x,s_0,k\sigma^2)$$

and, for case 2,

$$P_X(x) = \sum_k P_{X|K}(x|k)P_K(k) = \sum_k P_{S_k}(x)P_K(k) = \frac{1}{L}\sum_k P_{S_k}(x).$$

Noting that

$$P_{S_k}(x) = \int P_{S_k|S_0}(x|s_0)P_{S_0}(s_0)ds_0 = \int G(x,s_0,k\sigma^2)G(s_0,\mu,\nu^2)ds_0 = G(x,\mu,k\sigma^2+\nu^2)$$

we have, in case 2,

$$P_X(x) = \frac{1}{L}\sum_k G(x,\mu,k\sigma^2+\nu^2).$$

Hence, in both cases, we have a mixture of the form

$$P_X(x) = \frac{1}{L}\sum_k G(x,\alpha,k\sigma^2+\beta).$$

where α and β are known. The ML estimation is performed with EM. Using the standard trick of introducing binary vectors $\mathbf{z}_i$ to represent the assignments, the complete data log-likelihood is

$$\log P_{X,Z}(\mathcal{D},\{\mathbf{z}_1,\ldots,\mathbf{z}_n\}) = \sum_{i,j} z_{i,j}\log G(x_i,\alpha,j\sigma^2+\beta).$$

As usual, in the E-step we estimate the posterior assignment probabilities

$$h_{ij} = \frac{G(x_i, \alpha, j\sigma^2 + \beta)}{\sum_l G(x_i, \alpha, l\sigma^2 + \beta)}$$

and in the M-step we maximize

$$(\sigma^2)^* \quad = \quad \arg\max_{\sigma^2} -\sum_{i,j} h_{i,j} \frac{(x_i - \alpha)^2}{2(j\sigma^2 + \beta)} - \sum_{i,j} h_{i,j} \frac{1}{2} \log(j\sigma^2 + \beta).$$

Setting derivatives to zero, we get

$$\frac{dQ}{d\sigma^2} = \sum_{i,j} h_{i,j} \frac{j(x_i - \alpha)^2}{2(j\sigma^2 + \beta)^2} - \sum_{i,j} h_{i,j} \frac{j}{2(j\sigma^2 + \beta)} = 0.$$

or

$$\sum_j \frac{j \sum_i h_{i,j} (x_i - \alpha)^2}{2(j\sigma^2 + \beta)^2} - \sum_j \frac{j \sum_i h_{i,j}}{2(j\sigma^2 + \beta)} = 0$$

$$\sum_j \frac{j\xi_j}{(j\sigma^2 + \beta)^2} = \sum_j \frac{j\epsilon_j}{(j\sigma^2 + \beta)}$$

When $\beta = 0$ this leads to

$$\sum_j \frac{\xi_j}{j} = n\sigma^2$$

or

$$\sigma^2 = \frac{1}{n} \sum_j \frac{1}{j} \sum_i h_{i,j} (x_i - \alpha)^2$$

but for $\beta \neq 0$ there is no closed form solution for σ^2. We would have to use a numerical optimization procedure.