Relevance vector machines (RVMs) {Lecture 16

SVM drawbacks: not probabilistic, hard to generalize to $K>2$, nuisance prms C, ϵ.

[RVMs are a Bayesian technique for regression & classification which yields sparse models.]

Regression RVMs

Recall that in regression,

$$p(t|\vec{x}, \vec{w}, \beta) = \mathcal{N}(t|\underbrace{y(\vec{x})}_{\text{"}\vec{w}^T\cdot\vec{\phi}(\vec{x})\text{ [bias included]}}, \underbrace{\beta^{-1}}_{=\sigma^2})$$

In RVMs, we use

$$y(\vec{x}) = \sum_{n=1}^{N} w_n k(\vec{x}, \vec{x}_n) + b$$

$M = N+1$ prms in total

similar to $y(\vec{x}) = \sum_n (a_n - \hat{a}_n) k(\vec{x}, \vec{x}_n) + b$ in SVMs

Training data:

$$X = \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_N \end{pmatrix}$$

$$\vec{t} = \underline{t_1, t_2, \ldots, t_N}$$

Then $\mathcal{L} = p(\vec{t}|X, \vec{w}, b) = \prod_{n=1}^{N} p(t_n|\vec{x}_n, \vec{w}, \beta^{-1})$.

Introduce $\underbrace{p(\vec{w}|\vec{\alpha})}_{\text{prior}} = \prod_{i=1}^{M} \mathcal{N}(w_i|0, \alpha_i^{-1})$

separate α_i for each w_i

$\vec{\alpha} = \overline{\alpha_1 \dots \alpha_M}$

Then $p(\vec{w}|\vec{t}, X, \vec{\alpha}, \beta) = \mathcal{N}(\vec{w}|\vec{m}, \Sigma)$ ← posterior for $\vec{w}$

$$\begin{cases} \vec{m} = \beta \Sigma \Phi^T \vec{t}, \\ \Sigma^{-1} = A + \beta \Phi^T \Phi \end{cases} \Leftarrow (3.53), (3.54)$$

$$A = \begin{pmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_M \end{pmatrix}$$

For RVMs, $\Phi \Rightarrow \underbrace{K}_{(N+1)\times(N+1)}$, with elements $k(\vec{x}_n, \vec{x}_m)$ kernel matrix [including the bias]

$$y(\vec{x}) = \sum_{n=1}^{N} w_n k(\vec{x}, \vec{x}_n) + \underbrace{b}_{w_{N+1}\, k(\vec{x}, \vec{x}_{N+1}),\ k = 1, \forall \vec{x}} = \sum_{n=1}^{N+1} w_n k(\vec{x}, \vec{x}_n)$$

Evidence approximation:

maximize $p(\vec{t}|X, \vec{\alpha}, \beta) = \int \underbrace{p(\vec{t}|X, \vec{w}, \beta)}_{\text{likelihood, gaussian}} \underbrace{p(\vec{w}|\vec{\alpha})}_{\text{prior, gaussian}} d\vec{w}$

$$p(\vec{t}|X, \vec{\alpha}, \beta) = \mathcal{N}(\vec{t}|\vec{0}, C), \text{ where}$$

$$\underbrace{C}_{N\times N} = \beta^{-1} \mathbb{I} + \Phi A^{-1} \Phi^T$$

More explicitly,

$$\log p(\vec{t}|X,\vec{\alpha},\beta) = \frac{N}{2}\log\beta + \frac{1}{2}\sum_{i=1}^{M}\log\alpha_i -$$

$$- \frac{1}{2}\vec{t}^{\,T}C^{-1}\vec{t} + \frac{1}{2}\log|\Sigma| - \frac{N}{2}\log(2\pi)$$

$$\Uparrow$$

$$\log|C| = \log|\Sigma| - N\log\beta - \sum_{i=1}^{M}\log\alpha_i$$

[see next page]

Then, first of all

$$\frac{1}{2}\vec{t}^{\,T}C^{-1}\vec{t} = \frac{1}{2}\vec{t}^{\,T}(\beta^{-1}\mathbb{I} + \Phi A^{-1}\Phi^T)^{-1}\vec{t} \quad (=)$$

[Woodbury identity:

$$(\underbrace{A}_{\beta^{-1}\mathbb{I}} + \overset{\Phi}{\overbrace{B}}\underbrace{D^{-1}}_{A^{-1}}\underbrace{C}_{\Phi^T})^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}$$]

$$(=)\ \frac{1}{2}\vec{t}^{\,T}(\beta\mathbb{I} - \beta\Phi\underbrace{(A+\beta\Phi^T\Phi)^{-1}}_{\Sigma}\Phi^T\beta)\vec{t} =$$

$$= \frac{1}{2}(\beta\vec{t}^{\,T}\vec{t} - \underbrace{\beta\vec{t}^{\,T}\Phi\Sigma}_{\vec{m}^T}\underbrace{\Sigma^{-1}\Sigma}\underbrace{\Phi^T\vec{t}\beta}_{\vec{m}}) =$$

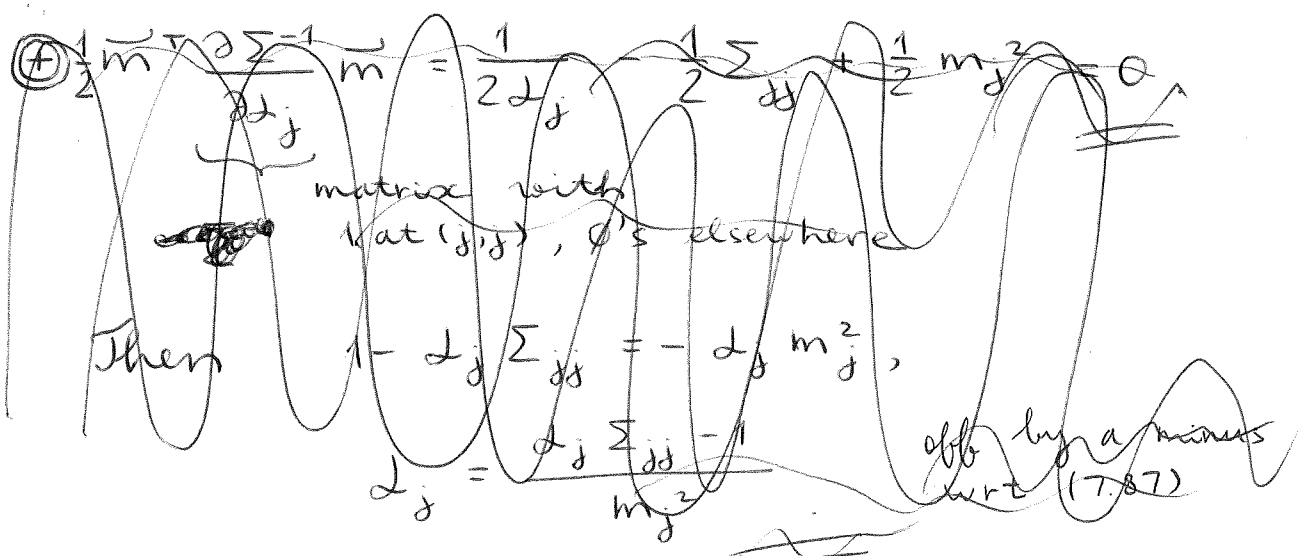
$$= \frac{1}{2}(\beta\vec{t}^{\,T}\vec{t} - \vec{m}^T\Sigma^{-1}\vec{m})$$

Next, $\frac{\partial}{\partial\alpha_j}\log p(\vec{t}|X,\vec{\alpha},\beta) = \frac{1}{2\alpha_j} + \frac{1}{2}\underbrace{\frac{\partial}{\partial\alpha_j}\log|\Sigma|}$ (+)

$$\overset{\Sigma^{-1}}{\overbrace{\quad}}\ \mathrm{Tr}\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\alpha_j}\right) \boxplus$$

$$\boxplus\ -\mathrm{Tr}\left(\Sigma^{-1}\Sigma\underbrace{\frac{\partial}{\partial\alpha_j}(A+\beta\Phi^T\Phi)}\Sigma\right) = -\Sigma_{jj}$$

$\mathbb{I}_{(j,j)}$: matrix with 1 at (j,j), 0's everywhere else



Double-check $\log|C|$:

$$|C| = |\beta^{-1}\mathbb{I}_N + \varphi A^{-1}\varphi^T| = |(\beta^{-1}\mathbb{I}_N)(\mathbb{I}_N + \beta\varphi A^{-1}\varphi^T)| =$$

$$= |\beta^{-1}\mathbb{I}_N|\,|\mathbb{I}_N + \beta\varphi A^{-1}\varphi^T| = \beta^{-N}|\mathbb{I}_M + \beta A^{-1}\varphi^T\varphi| =$$

$|AB| = |A||B|$; $|\mathbb{I}_N + AB^T| = |\mathbb{I}_M + A^TB|$; "A" "B^T"

$$= \beta^{-N}|A^{-1}(A + \beta\varphi^T\varphi)| = \beta^{-N}|A^{-1}|\,|\Sigma^{-1}| \;(=)$$

$$|A^{-1}| = \frac{1}{\prod_i \alpha_i}$$

$$(=)\ \frac{1}{\beta^N \prod_i \alpha_i}\ \frac{1}{|\Sigma|}, \text{ so that}$$

$$\log|C| = -N\log\beta - \sum_{i=1}^{M}\log\alpha_i - \log|\Sigma|,$$

just as before

$$\oplus \quad \frac{1}{2}\frac{\partial \vec{m}^T}{\partial \alpha_j}\Sigma^{-1}\vec{m} + \frac{1}{2}\vec{m}^T\underbrace{\frac{\partial \Sigma^{-1}}{\partial \alpha_j}}_{\mathbb{I}_{(j,j)}}\vec{m} + \frac{1}{2}\vec{m}^T\Sigma^{-1}\frac{\partial \vec{m}}{\partial \alpha_j} \quad \diamondsuit$$

Note that

$$\frac{\partial \vec{m}}{\partial \alpha_j} = \beta\frac{\partial(\Sigma^{-1})^{-1}}{\partial \alpha_j}\varphi^T\vec{t} = -\beta\Sigma\underbrace{\frac{\partial(\Sigma^{-1})}{\partial \alpha_j}}_{\mathbb{I}_{(j,j)}}\Sigma\varphi^T\vec{t}\,, \text{ s.t.}$$

$$\frac{1}{2}\vec{m}^T\Sigma^{-1}\frac{\partial \vec{m}}{\partial \alpha_j} = -\frac{1}{2}\vec{m}^T\Sigma^{-1}\Sigma\,\mathbb{I}_{(j,j)}\underbrace{(\beta\Sigma\varphi^T\vec{t})}_{\vec{m}} =$$

$$= -\frac{1}{2}\vec{m}^T\mathbb{I}_{(j,j)}\vec{m} = -\frac{1}{2}m_j^2 .$$

$$\diamondsuit \quad \frac{1}{2\alpha_j} - \frac{1}{2}\Sigma_{jj} \bullet - \frac{1}{2}m_j^2 + \frac{1}{2}m_j^2 - \frac{1}{2}m_j^2 =$$

$$= \frac{1}{2\alpha_j} - \frac{1}{2}\Sigma_{jj} - \frac{1}{2}m_j^2 = 0 .$$

Then $1 - \alpha_j\Sigma_{jj} = \alpha_j m_j^2$, or

$$(*) \qquad \alpha_j = \frac{\overbrace{1-\alpha_j\Sigma_{jj}}^{\equiv\gamma_j}}{m_j^2} \Leftarrow (7.87)$$

Use (*) as an update eq'n:

$$\alpha_j^{new} = \underbrace{\frac{\gamma_j}{m_j^2}}_{\text{evaluated at } \vec{\alpha}^{old},\ \beta^{old}}$$

Similarly,

$$\frac{\partial}{\partial \beta} \log p(\vec{t} \mid X, \vec{\alpha}, \beta) = 0 \quad \text{yields (proof left as an exercise)}$$

$$\frac{1}{\beta^{new}} = \underbrace{\frac{\|\vec{t} - \Phi \vec{m}\|^2}{N - \sum_{i=1}^{M} \gamma_i}}_{\text{evaluated at } \vec{\alpha}^{old},\ \beta^{old}} \Leftarrow (7.88)$$

Algorithm: choose initial $(\vec{\alpha}, \beta)$

$\Downarrow$

evaluate $(\vec{m}, \Sigma)$

$\Downarrow$

update $(\vec{\alpha}, \beta)$ ↑

[Stop upon convergence, record $(\hat{\vec{\alpha}}, \hat{\beta})$]

It turns out that many α_i's are driven to very large values, so that posteriors for the corresponding w_i's are peaked at $\emptyset$ (with variance $\emptyset$), and thus the $\varphi_i(\vec{x})$ are removed from the model.

If we use $y(\vec{x}) = \sum_{n=1}^{N} w_n k(\vec{x}, \vec{x}_n) + b$,

the datapoints $\vec{x}_n$ for which $w_n \neq 0$ are called relevance vectors.

Thus sparse models are created automatically.

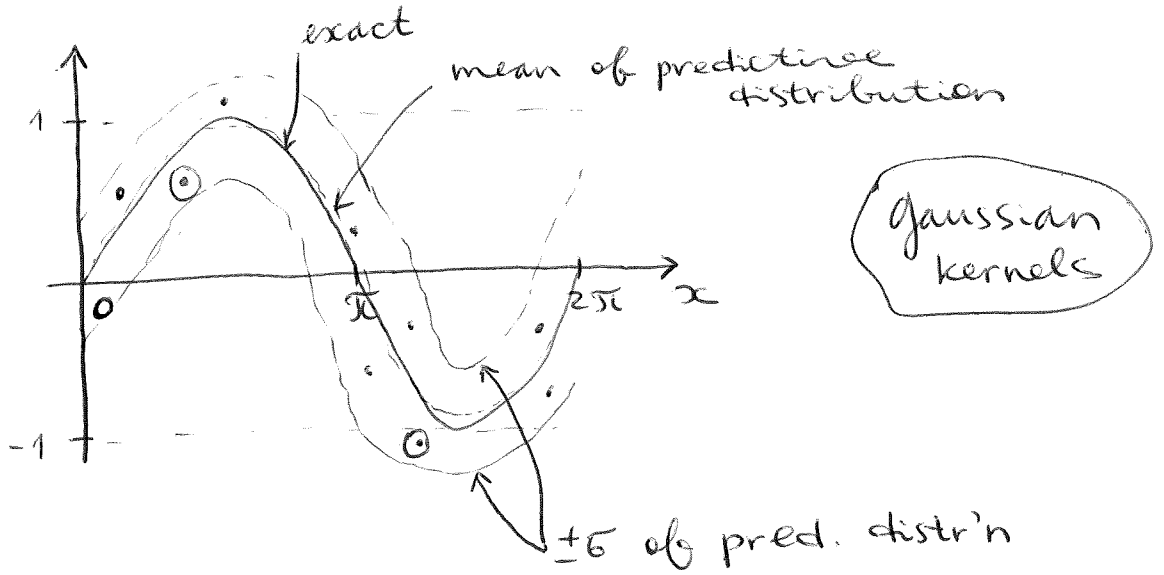
Finally,

$$p(t \mid \vec{x}, X, \vec{t}, \hat{\vec{\alpha}}, \hat{\beta}) = \int d\vec{w}\ p(t \mid \vec{x}, \vec{w}, \hat{\beta})\, p(\vec{w} \mid X, \vec{t}, \hat{\vec{\alpha}}, \hat{\beta}) \quad (11)$$

$$= \mathcal{N}(t \mid \vec{m}^T \vec{\varphi}(\vec{x}), \sigma^2(\vec{x})), \text{ where}$$

$$\sigma^2(\vec{x}) = \frac{1}{\hat{\beta}} + \vec{\varphi}(\vec{x})^T \Sigma\, \vec{\varphi}(\vec{x}) \quad \leftarrow \text{analogous to (3.58), (3.59)}$$

Ex:



⊙ are relevance vectors, there are relatively few.

Sparsity analysis

Ex.: consider $N=2$, s.t. we have $\vec{t}=\underline{t_1\ t_2}$, and a single basis f'n $\varphi(\vec{x})$ $(M=1)$ $\Rightarrow$ single α. Measurement noise is described by β.

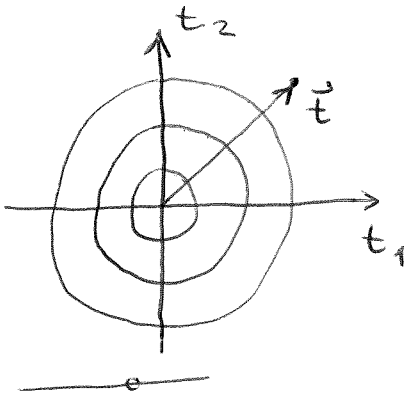
Then $p(\vec{t}\,|\,\alpha,\beta) = \mathcal{N}(\vec{t}\,|\,\vec{0}, C)$, where

$$\underbrace{C}_{2\times 2} = \beta^{-1}\mathbb{I} + \alpha^{-1}\vec{\varphi}\cdot\vec{\varphi}^T \quad \Leftarrow \quad \text{in general, } \underbrace{C}_{N\times N} = \beta^{-1}\mathbb{I} + \Phi A^{-1}\Phi^T$$

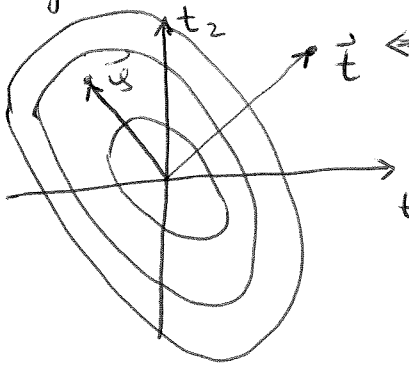
$$\vec{\varphi} = \underline{\varphi(\vec{x}_1),\ \varphi(\vec{x}_2)}$$

We maximize $p(\vec{t}\,|\,\alpha,\beta)$ to find $\hat{\alpha}, \hat{\beta}$.

$\alpha = \infty$: $C = \beta^{-1}\mathbb{I}$



α finite but $\vec{\varphi}$ poorly aligned with $\vec{t}$:



$\Leftarrow$ prob. of observing $\vec{t}$ decreases, so that $\alpha = \infty$ is preferable

Consider M basis functions.

Idea: make dependence on a specific α_i explicit in C & $\log p(\vec{t}\,|\,X, \vec{\alpha}, \beta) \equiv L(\vec{\alpha})$

$$C = \beta^{-1}\mathbb{I} + \alpha_i^{-1}\vec{\varphi}_i\vec{\varphi}_i^T + \sum_{j\neq i}\alpha_j^{-1}\vec{\varphi}_j\vec{\varphi}_j^T \equiv$$

$$\equiv \underbrace{C_{-i}}_{N\times N} + \alpha_i^{-1}\vec{\varphi}_i\cdot\vec{\varphi}_i^T$$

$$\vec{\varphi}_i = \underline{\varphi_i(\vec{x}_1) \ \dots \ \varphi_i(\vec{x}_N)}$$

$$C_{-i}^{-1} = (C_{-i}^{-1})^T$$

$$|C| = |C_{-i}|\,|\mathbb{I}_N + \underbrace{C_{-i}^{-1}\alpha_i^{-1}\vec{\varphi}_i}_{\text{"}\vec{a}\text{"}}\underbrace{\vec{\varphi}_i^T}_{\text{"}\vec{b}^T\text{"}}| = |C_{-i}|(1 + \alpha_i^{-1}\vec{\varphi}_i^T C_{-i}^{-1}\vec{\varphi}_i)$$

$$|\mathbb{I}_N + \underbrace{\vec{a}\cdot\vec{b}^T}_{\text{outer product}}| = 1 + \underbrace{\vec{a}^T\cdot\vec{b}}_{\text{inner product}}$$

$$C^{-1} = (\underbrace{C_{-i}}_{\text{"A"}} + \underbrace{\alpha_i^{-1}}_{\text{"}B\text{"}}\underbrace{\vec{\varphi}_i}_{B}\underbrace{\vec{\varphi}_i^T}_{\text{"C"}})^{-1} = C_{-i}^{-1} - C_{-i}^{-1}\vec{\varphi}_i(\underbrace{\alpha_i + \vec{\varphi}_i^T C_{-i}^{-1}\vec{\varphi}_i}_{\text{number}})^{-1} \times \vec{\varphi}_i^T C_{-i}^{-1} \equiv$$

(Woodbury identity)

$$\equiv C_{-i}^{-1} - \frac{C_{-i}^{-1}\vec{\varphi}_i\vec{\varphi}_i^T C_{-i}^{-1}}{\alpha_i + \vec{\varphi}_i^T C_{-i}^{-1}\vec{\varphi}_i}$$

Next, $L(\vec{\alpha}) = -\frac{1}{2}\vec{t}^T C^{-1}\vec{t} - \frac{N}{2}\log(2\pi) - \frac{1}{2}\log|C| =$

$$= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log\left\{|C_{-i}|(1+\alpha_i^{-1}\vec{\varphi}_i^T C_{-i}^{-1}\vec{\varphi}_i)\right\} -$$

$$-\frac{1}{2}\vec{t}^T\left\{C_{-i}^{-1} - \frac{C_{-i}^{-1}\vec{\varphi}_i\vec{\varphi}_i^T C_{-i}^{-1}}{\alpha_i + \vec{\varphi}_i^T C_{-i}^{-1}\vec{\varphi}_i}\right\}\vec{t} =$$

$$= \underbrace{-\frac{1}{2}\left\{N\log(2\pi) + \log|C_{-i}| + \vec{t}^T C_{-i}^{-1}\vec{t}\right\}}_{\text{"}L(\vec{\alpha}_{-i})\text{"}\,,\ \log\mathcal{Z}\ \text{with}\ \vec{\varphi}_i\ \text{omitted}} \ominus$$

$$\ominus \frac{1}{2}\log\left(1+\alpha_i^{-1}\vec{\varphi}_i^T C_{-i}^{-1}\vec{\varphi}_i\right) + \frac{1}{2}\vec{t}^T\frac{C_{-i}^{-1}\vec{\varphi}_i\vec{\varphi}_i^T C_{-i}^{-1}}{\alpha_i + \vec{\varphi}_i^T C_{-i}^{-1}\vec{\varphi}_i}\vec{t} =$$

$$= L(\vec{\alpha}_{-i}) \ominus \frac{1}{2}\log\left(1+\frac{s_i}{\alpha_i}\right) + \frac{1}{2}\frac{q_i^2}{\alpha_i+s_i} \quad (=)$$

$$\begin{cases} s_i = \vec{\varphi}_i^T C_{-i}^{-1}\vec{\varphi}_i, & \text{sparsity of } \vec{\varphi}_i \\ q_i = \vec{\varphi}_i^T C_{-i}^{-1}\vec{t} & \text{quality of } \vec{\varphi}_i \end{cases}$$

$$(=)\ L(\vec{\alpha}_{-i}) + \underbrace{\frac{1}{2}\left[\log\alpha_i - \log(\alpha_i+s_i) + \frac{q_i^2}{\alpha_i+s_i}\right]}_{\text{"}\lambda(\alpha_i)\text{"}\,,\ \text{contains all dependence on } \alpha_i}$$

Then $\frac{\partial L(\vec{\alpha})}{\partial\alpha_i} = \frac{d\lambda(\alpha_i)}{d\alpha_i} = \frac{1}{2}\left[\frac{1}{\alpha_i} - \frac{1}{\alpha_i+s_i} - \frac{q_i^2}{(\alpha_i+s_i)^2}\right] = \frac{1}{2}\left[\frac{(\alpha_i+s_i)^2 - \alpha_i(\alpha_i+s_i) - q_i^2\alpha_i}{(\alpha_i+s_i)^2\alpha_i}\right] =$

$$= \frac{1}{2}\frac{(\alpha_i^2+s_i^2+2\alpha_i s_i) - \alpha_i^2 - \alpha_i s_i - q_i^2\alpha_i}{(\alpha_i+s_i)^2\alpha_i} \quad \Diamond$$

$$\Leftrightarrow \frac{1}{2}\ \frac{s_i^2 + 2\alpha_i s_i - q_i^2 \alpha_i}{(\alpha_i + s_i)^2 \alpha_i} = \frac{\alpha_i^{-1} s_i^2 - (q_i^2 - s_i)}{2(\alpha_i + s_i)^2} = 0$$

If $q_i^2 \leq s_i$, only $\alpha_i \to \infty$ is a solution.

($\alpha_i \geq 0$)

If $q_i^2 > s_i$, $\underbrace{\alpha_i^{-1} s_i^2}_{>0} = \underbrace{q_i^2 - s_i}_{>0} \Rightarrow \alpha_i = \frac{s_i^2}{q_i^2 - s_i}$ (**)

for finite α_i

So, q_i^2 vs. s_i determine whether $\vec{\varphi}_i$ will be pruned from the model or not.

Furthermore, (**) gives α_i given $\vec{\alpha}_{-i}$. This gives rise to a sequential sparse Bayesian learning algorithm:

1. Initialize β
2. Set α_1 for $\vec{\varphi}_1$ using (**); set $\alpha_{j \neq 1}$ to ∞ so that only $\vec{\varphi}_1$ is included in the model

Repeat until convergence:

3. Compute Σ, $\vec{m}$, $\{q_i\}$, $\{s_i\}$.
4. Randomly select $\vec{\varphi}_i$.

→ If $q_i^2 > s_i$ & $\alpha_i < \infty$ ⇒ $\vec{\varphi}_i$ already in the model ⇒ update α_i using (**)

→ If $q_i^2 > s_i$ & $\alpha_i = \infty$ ⇒ add $\vec{\varphi}_i$ to the model, update α_i using (**)

→ if $q_i^2 \leq s_i$ & $\alpha_i < \infty$ ⇒ remove $\vec{\psi}_i$ from the model, set $\alpha_i = \infty$

→ if $q_i^2 \leq s_i$ & $\alpha_i = \infty$ ⇒ do nothing

5. Update β using (7.88) [see top of p. 6]

One can show that this algorithm scales as $O(M^3)$, where M is the number of active $\vec{\psi}_i$'s, typically $\ll N$.