Chapter 13. Sparse linear models

13.1.

a.

$$\text{RSS}(\mathbf{w}) = \sum_i (\mathbf{x}_i^T\mathbf{w} - y_i)^2 = \sum_i (\mathbf{x}_{i,-k}^T\mathbf{w}_{-k} + x_{ik}w_k - y_k)^2$$

$$\frac{\partial}{\partial w_k}\text{RSS}(\mathbf{w}) = \sum_i 2(\mathbf{x}_{i,-k}^T\mathbf{w}_{-k} + x_{ik}w_k - y_k)x_{ik} = a_k w_k - c_k.$$

b.

$$a_k w_k - c_k = 0 \Rightarrow \hat{w}_k = \frac{c_k}{a_k} = \frac{\mathbf{x}_{:,k}^T\mathbf{r}_k}{\|\mathbf{x}_{:,k}\|^2}.$$

13.2.

$$\frac{dQ'}{d\alpha_j} = \frac{d}{d\alpha_j}\left(\frac{1}{2}\sum_j \log\alpha_j - \frac{1}{2}\text{tr}(\mathbf{A}(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma})) + \sum_j (a\log\alpha_j - b\alpha_j)\right)$$

$$= \frac{1}{2\alpha_j} + \frac{a}{\alpha_j} - b - \frac{1}{2}(m_j^2 + \Sigma_{jj}) = 0$$

$$\Rightarrow \hat{\alpha}_j = \frac{1+2a}{m_j^2 + \Sigma_{jj} + 2b}.$$

$$\frac{dQ'}{d\beta} = \frac{N}{2\beta} - \frac{1}{2}(\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}\|^2 + \text{tr}(\mathbf{X}^T\mathbf{X}\boldsymbol{\Sigma})) + \frac{c}{\beta} - d.$$

By the hint, $\text{tr}(\mathbf{X}^T\mathbf{X}\boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{X}) = \beta^{-1}\sum_j(1 - \alpha_j\Sigma_{jj})$.

$$\Rightarrow \frac{dQ'}{d\beta} = 0$$

$$\Rightarrow \hat{\beta}^{-1} = \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}\|^2 + \beta^{-1}\sum_j (1 - \alpha_j \Sigma_{jj}) + 2d}{N + 2c}.$$

13.3.
Same with 13.2.

$$\alpha_j (m_j^2 + \Sigma_{jj} + 2b) = (1 + 2a) \Rightarrow \alpha_j (m_j^2 + 2b) = 2a + \gamma_j$$
$$\Rightarrow \alpha_j = \frac{2a + \gamma_j}{m_j^2 + 2b}.$$
$$\beta^{-1}(N + 2a) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}\|^2 + \beta^{-1}\sum_j \gamma_j + 2d$$
$$\Rightarrow \beta^{-1} = \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}\|^2 + 2d}{N + 2c - \sum_j \gamma_j}.$$

13.4.

$$p(\mathcal{D}|\boldsymbol{\gamma}) \propto |\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}|^{-\frac{1}{2}} |\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-\frac{1}{2}} (2b_\sigma + S(\boldsymbol{\gamma}))^{-\frac{2\alpha_\gamma + N - 1}{2}}$$
$$= |\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}} (1 + \frac{1}{g})|^{-\frac{1}{2}} |(\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})^{-1} g|^{-\frac{1}{2}} (2b_\sigma + S(\boldsymbol{\gamma}))^{-\frac{2\alpha_\gamma + N - 1}{2}}$$
$$= (1 + g)^{-\frac{D_{\boldsymbol{\gamma}}}{2}} (2b_\sigma + S(\boldsymbol{\gamma}))^{-\frac{2\alpha_\gamma + N - 1}{2}}.$$
$$S(\boldsymbol{\gamma}) = \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}_{\boldsymbol{\gamma}} (\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1})^{-1} \mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{y}$$
$$= \mathbf{y}^T \mathbf{y} - \frac{g}{1 + g} \mathbf{y}^T \mathbf{X}_{\boldsymbol{\gamma}} (\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})^{-1} \mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{y}.$$

13.5.

$$J_2(\mathbf{w}) = \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\tilde{\mathbf{w}}\|_2^2 + c\lambda_1 \|\tilde{\mathbf{w}}\|_1$$
$$= \tilde{\mathbf{y}}^T \tilde{\mathbf{y}} - 2\tilde{\mathbf{y}}^T \tilde{\mathbf{X}}\tilde{\mathbf{w}} + \tilde{\mathbf{w}}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}\tilde{\mathbf{w}} + c\lambda_1 \|\tilde{\mathbf{w}}\|_1$$
$$= \mathbf{y}^T \mathbf{y} - 2c\mathbf{y}^T \mathbf{X}\mathbf{w} + c^2(\mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w} + \lambda_2 \mathbf{w}^T \mathbf{w})$$
$$= \|\mathbf{y} - \mathbf{X}(c\mathbf{w})\|_2^2 + \lambda_2 \|c\mathbf{w}\|_2^2 + \lambda_1 \|c\mathbf{w}\|_1 = J_1(c\mathbf{w}).$$

13.6.
a. (1) OLS, (2) ridge, (3) lasso.

b. $\lambda_1 = 2$.
c. $\lambda_2 = 1$.

13.7.

$$p(\boldsymbol{\gamma}|\boldsymbol{\alpha}) = \int p(\boldsymbol{\gamma}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\boldsymbol{\alpha})d\boldsymbol{\pi}$$
$$= \prod_j [\int \mathrm{Ber}(\gamma_j|\pi_j)\mathrm{Beta}(\pi_j|\alpha_1, \alpha_2)d\pi_j]$$
$$= \prod_j \frac{B(\alpha_1 + \gamma_j, \alpha_2 + (1 - \gamma_j))}{B(\alpha_1, \alpha_2)} = \frac{\alpha_1^{\|\gamma\|_0}\alpha_2^{D-\|\gamma\|_0}}{(\alpha_1 + \alpha_2)^D}.$$

Using a prior on the sparsity obviously helps to encode prior knowledge into the distribution.
Using a prior on the sparsity has disadvantages that it is hard to decide the appropriate value of prior parameters.

13.8.

$$\mathbb{E}[\frac{1}{\tau_j^2}|w_j] = \int \frac{1}{\tau_j^2}p(\tau_j^2|w_j)d\tau_j^2 = \int \frac{1}{\tau_j^2}\frac{p(w_j|\tau_j^2)p(\tau_j^2)}{p(w_j)}d\tau_j^2$$
$$= \frac{1}{p(w_j)}\int \frac{1}{\tau_j^2}\mathcal{N}(w_j|0, \tau_j^2)p(\tau_j^2)d\tau_j^2 = \frac{1}{p(w_j)}\int -\frac{1}{|w_j|}\frac{d}{d|w_j|}\mathcal{N}(w_j|0, \tau_j^2)p(\tau_j^2)d\tau_j^2$$
$$= \frac{1}{|w_j|}\frac{1}{p(w_j)}\frac{dp(w_j)}{d|w_j|} = -\frac{1}{|w_j|}\frac{d\log p(w_j)}{d|w_j|} = \frac{\pi'(w_j)}{|w_j|}.$$

13.9.

$$\log p(\mathbf{w}|\mathbf{y}, \boldsymbol{\tau}, \mathbf{z}) \propto \log(p(\mathbf{z}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\tau})p(\boldsymbol{\tau})p(\mathbf{y}|\mathbf{z}))$$
$$= -\frac{1}{2}\|\mathbf{z} - \mathbf{X}\mathbf{w}\|_2^2 - \frac{1}{2}\mathbf{w}^T\boldsymbol{\Lambda}_{\boldsymbol{\tau}}\mathbf{w},$$

where $\boldsymbol{\Lambda}_{\boldsymbol{\tau}} = \mathrm{diag}(\frac{1}{\tau_j^2})$.
By Exercise 11-15,

$$\mathbb{E}[z_i|\mathbf{w}_{t-1}, \mathbf{x}_i] = \mu_i + \frac{\pi(\mu_i)}{\Phi(\mu_i)}$$

for $y_i = 1$,

$$\mathbb{E}[z_i|\mathbf{w}_{t-1}, \mathbf{x}_i] = \mu_i - \frac{\pi(\mu_i)}{1 - \Phi(\mu_i)}$$

for $y_i = 0$.
Denote this by s_i, and $\bar{\mathbf{W}}_{t-1} = \text{diag}(|w_{j,t-1}|)$.
Then since $\mathbb{E}[\mathbf{\Lambda}_{\boldsymbol{\tau}}^{-1}|\mathbf{w}] = \bar{\mathbf{W}}_{t-1}$,

$$Q(\mathbf{w}_t, \mathbf{w}_{t-1}) = -\frac{1}{2}\|\mathbf{s} - \mathbf{X}\mathbf{w}\|_2^2 - \frac{1}{2}\mathbf{w}^T\bar{\mathbf{W}}_{t-1}\mathbf{w},$$

$$\frac{dQ}{d\mathbf{w}_t} = 0 \Rightarrow \mathbf{w}_t = (\bar{\mathbf{W}}_{t-1} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{s}.$$

13.10.
If $\delta_g = \frac{d_g+1}{2}$,

$$p(\mathbf{w}_g) \propto |u_g|^{\frac{1}{2}}\sqrt{\frac{\pi}{2\rho u_g}}e^{-\rho u_g} \propto e^{-\gamma\|\mathbf{w}_g\|_2}.$$

$$\Rightarrow p(\mathbf{w}) \propto e^{-\gamma\sum_g\|\mathbf{w}_g\|_2}.$$

13.11.

$$\begin{aligned}\text{RSS}(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_1 &= \frac{1}{2}\|\mathbf{y} - \mathbf{X}(\boldsymbol{\theta}_+ - \boldsymbol{\theta}_-)\|_2^2 + \lambda\mathbf{1}^T\boldsymbol{\theta}_+ + \lambda\mathbf{1}^T\boldsymbol{\theta}_- \\ &= \frac{1}{2}\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\theta}_+ + \mathbf{y}^T\mathbf{X}\boldsymbol{\theta}_- + \frac{1}{2}(\boldsymbol{\theta}_+^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}_+^T - \boldsymbol{\theta}_+^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}_-^T \\ &\quad -\boldsymbol{\theta}_-^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}_+^T + \boldsymbol{\theta}_-^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}_-^T) + \lambda\mathbf{1}^T\boldsymbol{\theta}_+ + \lambda\mathbf{1}^T\boldsymbol{\theta}_- \\ &= \frac{1}{2}\mathbf{y}^T\mathbf{y} + \mathbf{c}^T\mathbf{z} + \frac{1}{2}\mathbf{z}^T\mathbf{A}\mathbf{z},\end{aligned}$$

where $\mathbf{c} = [\lambda\mathbf{1} - \mathbf{X}^T\mathbf{y}; \lambda\mathbf{1} + \mathbf{X}^T\mathbf{y}]$, $\mathbf{A} = [\mathbf{X}^T\mathbf{X}, -\mathbf{X}^T\mathbf{X}; -\mathbf{X}^T\mathbf{X}, \mathbf{X}^T\mathbf{X}]$,
$\mathbf{z} = [\boldsymbol{\theta}_+; \boldsymbol{\theta}_-]$.
Thus we have the problem minimizing $f(\mathbf{z}) = \frac{1}{2}\mathbf{y}^T\mathbf{y} + \mathbf{c}^T\mathbf{z} + \frac{1}{2}\mathbf{z}^T\mathbf{A}\mathbf{z}$ where $\mathbf{z} \geq \mathbf{0}$.
The projected gradient descent algorithm would be as follows:
Define $g_{ki} = \min(z_{ki}, \alpha(c_i + \mathbf{a}_i\mathbf{z}_k))$ where α is the learning rate. Then we can update $\mathbf{z}_{k+1} = \mathbf{z}_k - \mathbf{g}_k$.

13.12.
$\partial f(0) = -1$, $\partial f(1) = [-1, 0]$, $\partial f(2) = 0$.

13.13.
For n-variate polynomial $p(\mathbf{x})$, denote

$$\mathbf{x}^{\mathbf{r}} = \prod_j x_j^{r_j}, \sum_{\mathbf{r}}^{\mathbf{l}} = \sum_{r_1}^{l_1} \cdots \sum_{r_n}^{l_n}, \binom{\mathbf{l}}{\mathbf{r}} = \prod_j \binom{l_j}{r_j}.$$

Then $p(\mathbf{x}) = \sum_{\mathbf{r}}^{\mathbf{l}} a_{\mathbf{r}} \mathbf{x}^{\mathbf{r}}$ can be represented as $p(\mathbf{x}) = \sum_{\mathbf{r}}^{\mathbf{l}} b_{\mathbf{r}} B_{\mathbf{r}}(\mathbf{x})$, where

$$b_{\mathbf{r}} = \sum_{\mathbf{i}}^{\mathbf{r}} \frac{\binom{\mathbf{r}}{\mathbf{i}}}{\binom{\mathbf{l}}{\mathbf{i}}} a_{\mathbf{i}}, B_{\mathbf{r}}(\mathbf{x}) = \binom{\mathbf{l}}{\mathbf{r}} \mathbf{x}^{\mathbf{r}} (\mathbf{1} - \mathbf{x})^{\mathbf{l}-\mathbf{r}}$$

are Bernstein coefficients and Bernstein polynomials respectively.
Let $m = \prod_i (l_i + 1)$ and $\mathbf{C}$ be a $m \times (n+1)$ matrix defined as $c_{rj} = \frac{r_j}{l_j}$ for $1 \le j \le n$ and $c_{r,n+1} = 1$. Let $\mathbf{b}$ be the vector of corresponding m Bernstein coefficients. Then the coefficients $\boldsymbol{\gamma}$ of the affine function can be obtained by $\mathbf{C}^T \mathbf{C} \boldsymbol{\gamma} = \mathbf{C}^T \mathbf{b}$, yielding the affine function $g^*(\mathbf{x}) = \sum_i \gamma_i x_i + \gamma_{n+1}$.
Let $\delta^+ = \max_{\mathbf{r}} [g^*(\frac{\mathbf{r}}{\mathbf{l}}) - b_{\mathbf{r}}]$, then $g(\mathbf{x}) = g^*(\mathbf{x}) - \delta^+$ is the valid affine lower bound.