Parametric Empirical Bayes with ARD for Linear Regression

Dr Alexander D. Shaw
a.d.shaw@exeter.ac.uk
https://cpnslab.com

1 Parametric Empirical Bayes with ARD for Linear Regression

Model (single output). Given N observations and p predictors, let $X \in \mathbb{R}^{N\times p}$ and $y \in \mathbb{R}^N$. We assume a Gaussian likelihood

$$p(y \mid \beta, \sigma^2) \;=\; \mathcal{N}\big(y;\; X\beta,\; \sigma^2 I_N\big), \tag{1}$$

and an ARD Gaussian prior over coefficients

$$p(\beta \mid \Lambda) \;=\; \mathcal{N}\big(\beta;\; 0,\; \Lambda^{-1}\big), \qquad \Lambda \;=\; \mathrm{diag}(\lambda_1, \ldots, \lambda_p),\;\; \lambda_j > 0. \tag{2}$$

Posterior over coefficients. The posterior is Gaussian

$$\Sigma \;=\; \Big(\Lambda + \tfrac{1}{\sigma^2} X^{\mathsf{T}} X\Big)^{-1}, \tag{3}$$

$$\mu \;=\; \tfrac{1}{\sigma^2}\, \Sigma X^{\mathsf{T}} y. \tag{4}$$

Type-II Maximum Likeihood / Evidence maximisation. Define the marginal likelihood (evidence)

$$\log p(y \mid \Lambda, \sigma^2) = -\tfrac{1}{2}\Big(N \log 2\pi + \log\big|C\big| + y^{\mathsf{T}} C^{-1} y\Big), \qquad C \;=\; \sigma^2 I_N + X\Lambda^{-1} X^{\mathsf{T}}. \tag{5}$$

Equivalently, combining posterior terms gives the standard free-energy form:

$$\mathcal{F}(\Lambda, \sigma^2) = \tfrac{1}{2}\Big(\log|\Lambda| + N\log\tfrac{1}{\sigma^2} - \log|A|\Big) - \tfrac{1}{2}\Big(\tfrac{1}{\sigma^2}\|y - X\mu\|^2 + \mu^{\mathsf{T}}\Lambda\mu + N\log 2\pi\Big), \quad A = \Lambda + \tfrac{1}{\sigma^2} X^{\mathsf{T}} X, \tag{6}$$

which is equal to $\log p(y \mid \Lambda, \sigma^2)$ up to constants. The standard EM/Type-II updates use the *effective degrees of freedom*

$$\gamma_j \;=\; 1 - \lambda_j\, \Sigma_{jj}, \qquad j = 1, \ldots, p. \tag{7}$$

Then

$$\lambda_j^{\text{new}} \;=\; \frac{\gamma_j}{\mu_j^2 + \varepsilon}, \qquad \varepsilon > 0 \text{ small (stability)}, \tag{8}$$

$$\sigma^{2\,\text{new}} \;=\; \frac{\|y - X\mu\|^2}{N - \sum_{j=1}^{p} \gamma_j}. \tag{9}$$

Equations (3)–(9) are iterated until convergence.

Predictive distribution. For a new design row $x_\star \in \mathbb{R}^p$,

$$p(y_\star \mid x_\star, \mathcal{D}) = \mathcal{N}\big(y_\star;\ x_\star^\mathsf{T}\mu,\ \sigma^2 + x_\star^\mathsf{T}\Sigma x_\star\big). \tag{10}$$

Multi-output extension (shared sparsity). For d outputs, $Y \in \mathbb{R}^{N\times d}$, collect coefficients in $B \in \mathbb{R}^{p\times d}$ with columns $b_{(k)}$. Use independent Gaussian likelihoods with per-output noise σ_k^2 and a shared ARD prior

$$p(B \mid \Lambda) \ = \ \prod_{k=1}^{d} \mathcal{N}\big(b_{(k)};\ 0,\ \Lambda^{-1}\big), \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p). \tag{11}$$

Posteriors (for each k):

$$\Sigma \ = \ \Big(\Lambda + \tfrac{1}{\sigma_k^2} X^\mathsf{T} X\Big)^{-1}, \qquad \mu_{(k)} \ = \ \tfrac{1}{\sigma_k^2}\,\Sigma X^\mathsf{T} y_{(k)}. \tag{12}$$

Share the ARD precisions via the common Σ (or its diagonal):

$$\gamma_j \ = \ 1 - \lambda_j\,\Sigma_{jj}, \tag{13}$$

$$\lambda_j^{\text{new}} \ = \ \frac{\gamma_j}{\frac{1}{d}\sum_{k=1}^{d} \mu_{j(k)}^2 + \varepsilon}, \tag{14}$$

$$\sigma_k^{2\,\text{new}} \ = \ \frac{\|y_{(k)} - X\mu_{(k)}\|^2}{N - \sum_{j=1}^{p} \gamma_j}, \qquad k = 1, \ldots, d. \tag{15}$$

Predictive covariance for output k:

$$\mathrm{Var}\big[y_{\star(k)} \mid x_\star\big] \ = \ \sigma_k^2 + x_\star^\mathsf{T}\Sigma x_\star. \tag{16}$$

Intercept and standardisation (as in code). If an intercept column is included in X, one may leave it unpenalised by fixing a very small precision (or excluding it from ARD). For numerical stability, we standardise:

$$\tilde{X}_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j}, \quad \tilde{y} = \frac{y - \bar{y}}{s_y},$$

fit in $(\tilde{X}, \tilde{y})$ to obtain $(\tilde{\mu}, \tilde{\Sigma})$, then map back:

$$\mu_j = \frac{s_y}{s_j}\,\tilde{\mu}_j, \qquad \Sigma_{jj'} = \frac{s_y^2}{s_j s_{j'}}\,\tilde{\Sigma}_{jj'}.$$

Algorithm 1 PEB–ARD (shared sparsity, multi-output)

Input: X, Y; initialise $\lambda_j \leftarrow 1$, $\sigma_k^2 \leftarrow \mathrm{Var}(y_{(k)})$.
repeat
  Form $A = \Lambda + \frac{1}{\bar{\sigma}^2} X^\mathsf{T} X$ with $\bar{\sigma}^2 = \frac{1}{d}\sum_k \sigma_k^2$.
  Compute $\Sigma = A^{-1}$, and $\mu_{(k)} = \frac{1}{\sigma_k^2}\Sigma X^\mathsf{T} y_{(k)}$ for all k.
  $\gamma_j \leftarrow 1 - \lambda_j \Sigma_{jj}$.
  $\lambda_j \leftarrow \dfrac{\gamma_j}{\frac{1}{d}\sum_k \mu_{j(k)}^2 + \varepsilon}$ (clip to $[\lambda_{\min}, \lambda_{\max}]$).
  $\sigma_k^2 \leftarrow \dfrac{\|y_{(k)} - X\mu_{(k)}\|^2}{N - \sum_j \gamma_j}$ (floor at $\sigma_{\min}^2$).
until converged

Algorithm (Type-II EM). In practice we compute Σ and solves with Cholesky plus jitter; if needed we apply an SVD-based SPD repair. We also prune rank-deficient columns of X via QR before fitting (always retaining a constant column if present).

Numerical implementation. To ensure numerical stability when $X^\mathsf{T}X$ or A is nearly singular, the implementation performs Cholesky factorisation with progressively increased *jitter* ϵI until positive definiteness is achieved. If Cholesky fails, the symmetric matrix $(A + A^\mathsf{T})/2$ is repaired via singular value decomposition (SVD) and eigenvalue flooring. Prior to fitting, the design matrix is orthogonalised with a rank-revealing QR decomposition to remove collinear or zero-variance columns, while always retaining any constant (intercept) column. All matrix inversions are avoided in favour of triangular solves.

Connection to empirical Bayes and ridge regression. This procedure maximises the marginal likelihood $p(Y \mid X, \Lambda, \sigma^2)$, treating (Λ, σ^2) as hyperparameters estimated at the group or "second" level—hence the term *parametric empirical Bayes (PEB)*. When all λ_j are constrained equal,

$$\lambda_j = \lambda \quad \forall j,$$

the update reduces to standard Bayesian ridge regression with precision λ and noise variance σ^2 estimated by evidence optimisation. Automatic relevance determination (ARD) generalises this by learning one precision per feature, leading to sparsity as irrelevant predictors acquire large λ_j.

Interpretation in variational-free-energy form. The negative free energy (evidence lower bound) for this Gaussian model is

$$\mathcal{F} = \mathbb{E}_q[\log p(y, \beta)] - \mathbb{E}_q[\log q(\beta)],$$

where $q(\beta) = \mathcal{N}(\mu, \Sigma)$ is the variational posterior. Under the conjugate Gaussian model, $\mathcal{F}$ is maximised exactly by the updates above, and therefore coincides with the log-evidence. In this view, the algorithm performs coordinate ascent on $\mathcal{F}$ with respect to $\{\mu, \Sigma, \Lambda, \sigma^2\}$.

Computational complexity. For p predictors and N samples, each iteration requires forming $X^\mathsf{T}X$ and $X^\mathsf{T}Y$ once ($\mathcal{O}(Np^2)$) and solving a $p \times p$ linear system ($\mathcal{O}(p^3)$) per update. In typical applications $p \ll N$, so runtime is dominated by the matrix multiplications. The QR pre-pruning reduces p to the numerical rank of X, ensuring stable and efficient computation.

Summary. The resulting posterior over coefficients

$$\beta \mid Y, X, \hat{\Lambda}, \hat{\sigma}^2 \;\sim\; \mathcal{N}\big(\mu, \Sigma\big)$$

provides both point estimates and credible intervals (from Σ) for inference, as well as predictive uncertainty

$$y_\star \mid x_\star, \mathcal{D} \;\sim\; \mathcal{N}\big(x_\star^\mathsf{T}\mu,\; \hat{\sigma}^2 + x_\star^\mathsf{T}\Sigma x_\star\big).$$

This compact form directly underlies the accompanying MATLAB routine `peb_ard_novar.m`, which implements the updates in Eqs. (3)–(9) using Cholesky/SVD solves, automatic relevance determination, and optional shared sparsity across multiple outputs.