1 Baseball batting example, revisited

We shortly revisit the famous baseball batting example (Efron and Morris, 1975), often used as a scholarly example of EB. While this is an estimation problem instead of a prediction problem, we revisit it for several reasons: i) it is a well-known example for which the true values are known; ii) the EB objective function is the same as for diagonal linear discriminant analysis; and iii) by casting the problem to a large \( p \) setting it allows us to show the importance of \( p \) being large.

For 18 baseball players, their batting averages over the first 45 bats are recorded and denoted by \( B_i \). The batting averages over the remainder of the season are also known, and considered to be the truth. We follow Van Houwelingen (2014) by modeling each player’s batting average \( B_i \) as a Gaussian random variable, \( B_i \sim N(\mu_i, \tau_i^2) \), with the aim to estimate \( \theta_i \). The variances are estimated by \( \hat{\sigma}_i^2 = B_i(1-B_i)/45 \). Then, to effectuate shrinkage Van Houwelingen (2014) applies a Gaussian prior \( N(\mu, \tau^2) \) to \( \theta_i \). In the formulation of the marginal likelihood (see Main Document), this implies hyper-parameter \( \alpha = (\mu, \tau^2) \), and estimation of \( \alpha \) is straightforward due to the conjugacy of the likelihood and the prior: \( \hat{\tau}^2 = V(B_i) - \hat{\sigma}_i^2 \) and \( \hat{\mu} = \sum_{i=1}^n w_i B_i / \sum_{i=1}^n w_i \), with \( w_i = (\hat{\tau}^2 + \hat{\sigma}_i^2)^{-1} \). Then, the posterior mean estimate equals \( \hat{\theta}_i = E(\theta_i|B_i; (\hat{\mu}, \hat{\tau}^2)) = \hat{\mu} + \hat{\tau}^2 (\hat{\tau}^2 + \hat{\sigma}_i^2)^{-1} (B_i - \hat{\mu}) \). The conclusion in Van Houwelingen (2014) is that the shrinkage prior slightly reduces the mean squared error, but enforces too strong shrinkage for the extremes. E.g. for the best player \( \hat{\theta}_1 = 0.271 \), whereas \( X_1 = 0.400 \) and true \( \theta_1 = 0.346 \). Two possible explanations come to mind: the estimate of the prior parameters is not good due to \( p = n \) being small and/or the prior does not accommodate the extremes well. We investigate this.

First, we simulate 10,000 additional true values from a density estimate with Gaussian kernel (using R’s \texttt{density} command) applied to \((\theta_1, \ldots, \theta_{18})\). To obtain \( B_i \), \( i = 19, \ldots, 10018 \), Gaussian noise was added with variances \( \theta_i(1-\theta_i)/45 \). The estimates obtained in Van Houwelingen (2014) were \( \hat{\mu} = 0.256 \) and \( \hat{\tau}^2 = 0.000623 \). The latter seems to be a major cause of over-shrinkage: the true variance computed from the 18 known \( \theta_i \)’s equals 0.00143. If we estimate \( \tau^2 \) from the large data set, a much better estimate is obtained: \( \hat{\tau}^2 = 0.00195 \), as compared to the variance of the 18 known plus 10,000 generated true \( \theta_i \)'s, equaling 0.00166. From this, we obtain posterior mean estimate \( \hat{\theta}_1 = 0.293 \), which is substantially closer to \( \theta_1 = 0.346 \) than \( \hat{\theta}_1 = 0.271 \). Estimates for all 18 players are displayed in Figure 1(a).

In this example, it is natural to replace the Gaussian prior by a 3-component Gaussian mixture prior (bad, mediocre and good players): \( \theta_i \sim \sum_{k=1}^3 p_k N(\mu_k, \tau_k^2) \). Then, \( \alpha \) consists of 8 hyperparameters given that \( p_3 = 1 - p_1 - p_2 \). We employed the EM-type algorithm of Van de Wiel et al. (2012) to maximize the marginal likelihood (see Main Document) in terms of \( \alpha \). Here, we
use that the likelihood is Gaussian, and the Gaussian mixture prior is conjugate to it. The latter also facilitates straightforward computation of the shrunken estimator $\hat{\theta}^{\text{Mixt}}_i = E(\theta|B_i; \alpha)$. In this setting, the mixture prior is fairly close to the estimated Gaussian prior, and so are the shrunken estimates, as displayed in Figure 1(b). Slightly less shrinkage for the extremes is observed, though. For example, $\hat{\theta}^{\text{Mixt}}_1 = 0.298$.

### 2 Bayesian elastic net

The Bayesian linear elastic net model, as used in the Main Document, is (Li and Lin, 2010):

$$Y|X, \beta, \sigma^2 \sim N(X\beta, \sigma^2 I)$$

$$\beta | \sigma^2 \sim p\prod_{j=1}^{p} \text{GIG}(1/2, \psi = \lambda_2 \sigma^2/\sqrt{4\lambda_2 \sigma^2}, \chi_j = \lambda_2 \beta_j^2/\sigma^2),$$

with some arbitrary (possibly improper) density $f(\sigma^2)$. The normalizing constant $g(\lambda_1, \lambda_2, \sigma^2)$ is given by:

$$g(\lambda_1, \lambda_2, \sigma^2) = \sqrt{\frac{\lambda_2}{4\sigma^2}} \phi\left(\frac{\lambda_1}{\sqrt{4\lambda_2 \sigma^2}}\right) \Phi\left(-\frac{\lambda_1}{\sqrt{4\lambda_2 \sigma^2}}\right)^{-1}$$

Since the simulations are for illustrative purposes only, the error variance was kept fixed at its true value ($\sigma^2 = 1$) throughout the simulations. Then, after introducing the latent variables $\tau = [\tau_1 \cdots \tau_p]^T$, we have the following conditional distributions for $\beta$ and $\tau$:

$$\beta|Y, \sigma^2, \tau \sim N(A^{-1}X^TY, \sigma^2 A^{-1})$$

$$(\tau - 1)|Y, \sigma^2, \beta \sim \prod_{j=1}^{p} \text{GIG}\left(1/2, \psi = \lambda_2 \sigma^2/\sqrt{4\lambda_2 \sigma^2}, \chi_j = \lambda_2 \beta_j^2/\sigma^2\right),$$

where $A = X^TX + \lambda_2 \text{diag}[\tau_j/((\tau_j - 1))]$, and GIG denotes the generalized inverse Gaussian distribution.

### 2.1 Marginal likelihood from Gibbs samples

According to Chib (1995), the log marginal likelihood of a Bayesian model may be calculated from the converged Gibbs samples as:

$$\log m(Y) = \log \left[\frac{f(Y|\beta^*) p(\beta^*)}{p(\beta^*|Y)}\right] \approx \log \left[\frac{1}{K} \sum_{k=1}^{K} \frac{f(Y|\beta^{(k)})}{p(\beta^{(k)}|Y)}\right], \quad (1)$$

where $\beta^*$ is some high posterior density point of $p(\beta)$ and $\tau^{(k)}$ are Gibbs samples indexed by $k = 1, \ldots, K$. In principle, any point $\beta^*$ may be used, but for the sake of efficiency a high-density point of $\beta$ is preferred, such as the posterior mode. Then, for fixed $\sigma^2$, the log marginal likelihood is approximated by:
log \(m(Y) \approx \log k - \log \left\{ \sum_{k=1}^{K} \exp \left( \frac{n - p}{2} \log(2\pi) + \frac{n}{2} \log \sigma^2 + \frac{p}{2} \log 4 - \frac{n}{2} \log \lambda_2 \right) - p \log \phi \left( \frac{\lambda_1}{\sqrt{4\lambda_2\sigma^2}} \right) + p \log \Phi \left( \frac{-\lambda_1}{\sqrt{4\lambda_2\sigma^2}} \right) + \frac{\lambda_1}{2\sigma^2} \sum_{j=1}^{p} \beta_j^2 - \frac{\lambda_2}{2\sigma^2} \sum_{j=1}^{p} \frac{(\beta_j^*)^2}{\tau_j^{(k)}} \right\} \),

where \(T^{(k)} = \text{diag}(\tau_j^{(k)})\).

Sampling from the multivariate normal is a costly operation in high dimensions. In Bhattacharya et al. (2015) an efficient sampling scheme for \(\beta\) is described:

Set \(D = \lambda_2^{-1}I_p + \lambda_2^{-1}\text{diag}(\tau_j^{-1})\)

Generate \(u \sim N(0,D)\)

Generate \(v \sim N(0,I_n)\)

\(\text{return } \beta = u + DX^T(XDX^T + I_n)^{-1}(Y - Xu + 1)\).

Furthermore, if \((\tau_j - 1)|Y, \sigma^2, \beta_j \sim \text{GIG}(1/2, \psi, \chi_j)\), then \(1/(\tau_j - 1)|Y, \sigma^2, \beta_j \sim \text{IGauss}(\mu_j = \sqrt{\psi/\chi_j}, \lambda = \psi)\). Sampling from this inverse Gaussian is done by the following scheme:

Generate \(U \sim U(0,1)\)

Generate \(Y \sim N(0,1)\)

Set \(z = \sqrt{\psi/\chi_j} + y^2/2\chi_j - \sqrt{\psi y^2/4\chi_j^3} + y^4/4\chi_j^5\)

if \(u \leq (1 + z\sqrt{\chi_j/\psi})^{-1}\) then

\(\text{return } \tau_j = 1/z + 1\)

else

\(\text{return } \tau_j = \chi_j z/\psi + 1\)

end if

3 Proof for Theorem 1: EMSE \(\tau^2\) for linear regression

First, write

\[ \text{EMSE}(\hat{\tau}^2) = E_\beta\left[ \text{MSE}(\hat{\tau}^2|\beta, X) \right] = E_\beta\left[ E_\beta\left[ (E_Y(\hat{\tau}^2|\beta, X) - \tau^2)^2 + V_Y(\hat{\tau}^2|\beta, X) \right] \right]. \] (2)
Then, let us first compute the expected squared bias w.r.t. $\beta$:

$$E_\beta(\text{bias}^2) = E_\beta[(E_Y(\hat{\tau}^2|\beta, X) - \tau^2)^2] = p^{-2}E_\beta[(\sum_{j=1}^{p}(E_Y(\hat{\beta}_j^2) - v_j) - p\tau^2)^2]$$

$$= p^{-2}E_\beta[(\sum_{j=1}^{p}\beta_j^2 - p\tau^2)^2] = p^{-2}E_\beta[\sum_{j=1}^{p}\beta_j^2 - 2p\tau^2\sum_{j=1}^{p}\beta_j^2 + p^2\tau^4]$$

$$= p^{-2}(3\tau^4 + 2\sum_{j,k\neq j}\tau^4 - p^2\tau^4) = p^{-2}(3p\tau^4 + (p-1)\tau^4 - p^2\tau^4)$$

$$= \frac{2\tau^4}{p},$$

where we used the central moments of Gaussian random variables, available from Isserlis’ Theorem (Isserlis, 1918): $E(\hat{\beta}_j^4) = 3\tau^4$ and $E(\hat{\beta}_j^2\hat{\beta}_k^2) = E(\hat{\beta}_j^2)E(\hat{\beta}_k^2) = \tau^4$. The result is constant in $X$, so

$$E_X[E_\beta(\text{bias}^2)] = \frac{2\tau^4}{p}. \tag{3}$$

Next, we compute $E_X[E_\beta[V_Y(\hat{\tau}^2|\beta, X)]].$ Denoting $V(\hat{\tau}^2) = V_Y(\hat{\tau}^2|\beta, X)$, we have:

$$V(\hat{\tau}^2) = V((\hat{\tau}')^2) = \frac{1}{p^2}\left(\sum_{j=1}^{p}V(\hat{\beta}_j^2) + \sum_{j,k\neq j}\text{Cov}(\hat{\beta}_j^2, \hat{\beta}_k^2)\right). \tag{4}$$

Hence, we need to compute $V(\hat{\beta}_j^2)$ and $\text{Cov}(\hat{\beta}_j^2, \hat{\beta}_k^2)$. These are again derived from expressions of the central moments of Gaussian random variables. Let us first express the non-central moments in $\text{Cov}(\hat{\beta}_j^2, \hat{\beta}_k^2) = \text{Cov}(\hat{\beta}_j^2, \hat{\beta}_k^2) = E(\hat{\beta}_j^2\hat{\beta}_k^2) - E(\hat{\beta}_j^2)E(\hat{\beta}_k^2)$ in terms of the central ones. Denote the centralized value of $\hat{\beta}_j$ by $\hat{\beta}_j - \hat{\beta}_j$. Then,

$$E[\hat{\beta}_j^2\hat{\beta}_k^2] = E[(\hat{\beta}_j - \hat{\beta}_j + \hat{\beta}_j)(\hat{\beta}_k - \hat{\beta}_k + \hat{\beta}_k)^2] = E[(\hat{\beta}_j + \hat{\beta}_j)(\hat{\beta}_k + \hat{\beta}_k)^2]$$

$$= T_1 + T_2 := E[\hat{\beta}_j^2\hat{\beta}_k^2 + 4\hat{\beta}_j\hat{\beta}_k\hat{\beta}_j\hat{\beta}_k + \hat{\beta}_j^2\hat{\beta}_k^2 + \hat{\beta}_j^2\hat{\beta}_k^2 + \hat{\beta}_j^2\hat{\beta}_k^2]$$

$$+ E[2\hat{\beta}_j\hat{\beta}_j\hat{\beta}_k^2 + 2\hat{\beta}_j\hat{\beta}_k\hat{\beta}_j^2 + 2\hat{\beta}_j\hat{\beta}_j\hat{\beta}_k^2 + 2\beta_k\hat{\beta}_j\beta_j^2]$$

$$= T_1,$$

because $T_2 = 0$ due to the symmetry of the central Gaussian distribution. Likewise, the second term of the covariance equals:

$$E(\hat{\beta}_j^2)E(\hat{\beta}_k^2) = E(\hat{\beta}_j^2)E(\hat{\beta}_k^2) + \beta_j^2E(\hat{\beta}_j^2) + \beta_k^2E(\hat{\beta}_k^2) + \beta_j^2\beta_k^2.$$

Subtracting the latter from $T_1$ cancels the latter 3 terms in both expressions, rendering

$$\text{Cov}(\hat{\beta}_j^2, \hat{\beta}_k^2) = E(\hat{\beta}_j^2\hat{\beta}_k^2) + 4\beta_j\beta_kE(\hat{\beta}_j\hat{\beta}_k) - E(\hat{\beta}_j^2)E(\hat{\beta}_k^2)$$

$$= (v_jv_k + 2v_j^2) + 4\beta_j\beta_kv_{jk} - v_jv_k = 2v_{jk} + 4\beta_j\beta_kv_{jk}, \tag{5}$$

where we used the equations for the central moments of Gaussian random variables (Isserlis, 1918). Noting that $V(\hat{\beta}_j^2) = \text{Cov}(\hat{\beta}_j^2, \hat{\beta}_j^2)$ we directly obtain

$$V(\hat{\beta}_j^2) = E(\hat{\beta}_j^2) + 4\beta_j\beta_kE(\hat{\beta}_j\hat{\beta}_k) - (E(\hat{\beta}_j^2))^2$$

$$= 3v_j^2 + 4\beta_j^2v_j - v_j^2 = 2v_j^2 + 4\beta_j^2v_j, \tag{6}$$

4
Note that the latter can also be obtained by writing \( \hat{\beta}_j^2 = v_j(\hat{\beta}_j/\sqrt{v_j})^2 = v_j(\beta_j')^2 \). Then \( \beta_j' \sim N(\beta_j/\sqrt{v_j}, 1) \), so \( (\beta_j')^2 \sim \chi^2(\nu = \beta_j^2/v_j, k = 1) \) with \( V((\beta_j')^2) = 2(k + 2\nu) = 2(1 + 2\beta_j^2/v_j) \).

Hence, indeed \( V(\hat{\beta}_j^2) = v_j^2 V((\beta_j')^2) = 2v_j^2 + 4\beta_j^2 v_j \).

Substituting (5) and (6) into (4) renders:

\[
V(\hat{\tau}^2) = \frac{2}{p^2} \left[ \sum_{j=1}^{p} (v_j^2 + 2\beta_j^2 v_j) + \sum_{j,k \neq j} (v_{jk}^2 + 2\beta_j\beta_k v_{jk}) \right].
\]

Taking expectation w.r.t. \( \beta \) gives:

\[
E_{\beta}[V(\hat{\tau}^2)] = \frac{2}{p^2} \left[ \sum_{j=1}^{p} (v_j^2 + 2\tau^2 v_j) + \sum_{j,k \neq j} v_{jk}^2 \right],
\]

because we assume i.i.d. central priors for \( \beta_j \). Now to compute

\[
E_{X}[E_{\beta}[V(\hat{\tau}^2)]] = \frac{2}{p^2} \left[ \sum_{j=1}^{p} (E_{X}(v_j^2) + 2\tau^2 E_{X}(v_j)) + \sum_{j,k \neq j} E_{X}(v_{jk}^2) \right]
\]

we need to know \( E_{X}(v_j^2), E_{X}(v_j) \) and \( E_{X}(v_{jk}^2) \), where \( v_j = v_{jj} \) and \( v_{jk} = V_{jk}, V = (X^T X)^{-1} \), with \( X_i \sim N(0, \Sigma = \Sigma_{p \times p}) \) and \( \Sigma_{jj} = 1 \). By definition, \( V \) follows an inverse-Wishart distribution: \( V \sim \mathcal{W}^{-1}(\Psi = \Sigma^{-1}, n) \). Hence, the requested moments are known (Press, 1982):

\[
E_{X}(v_j^2) = \psi_{jj}/(n - p - 1)
\]

\[
E_{X}(v_j) = \sqrt{\psi_{jj}/(n - p - 1)}
\]

\[
E_{X}(v_{jk}^2) = V_{jk} + (E_{X}(v_j))^2 = \frac{(n - p + 1)\psi_{jk}^2 + (n - p + 1)\psi_{jj}\psi_{kk} + \psi_{jk}^2}{(n - p)(n - p - 1)(n - p - 3)} + \frac{\psi_{jk}^2}{(n - p - 1)^2},
\]

where we assume \( p < n - 3 \). Substituting (8) into (7) and aggregating with the expected squared bias (3) finalizes the result:

\[
\text{EMSE}(\hat{\tau}^2) = \frac{2}{(n - p - 1)p^2} \left[ \sum_{j=1}^{p} \frac{2\psi_{jj}^2}{(n - p - 1)(n - p - 3)} + \frac{\psi_{jj}^2}{(n - p - 1)} + 2\tau^2 \sum_{j=1}^{p} \psi_{jj} \right]
\]

\[
+ \frac{p}{(n - p)(n - p - 1)(n - p - 3)} + \frac{1 + \tau^2}{n - p - 1} + 2\tau^2 \sum_{j,k \neq j} \psi_{jk} \left[ \frac{(n - p + 1)\psi_{jk}^2 + (n - p - 1)\psi_{jj}\psi_{kk} + \psi_{jk}^2}{(n - p - 1)} \right] + \frac{2\tau^4}{p}.
\]

This simplifies for independent \( X_i \), because then \( \psi_{jj} = 1 \) and \( \psi_{jk} = 0 \):

\[
\text{EMSE}_{\perp}(\hat{\tau}^2) = \frac{2}{(n - p - 1)p} \left[ \frac{2}{(n - p - 1)(n - p - 3)} + \frac{1}{n - p - 1} + 2\tau^2 \right]
\]

\[
+ \frac{p - 1}{(n - p)(n - p - 3)} + \frac{2\tau^4}{p}.
\]

4 Supporting Figures

4.1 Baseball Example

4.2 Simulation Example

Here, we show the results for all simulation settings presented in the Simulation example.

5
Figure 1: Shrinkage estimators vs true value. Left: Parameter estimates for 18 real players when EB estimation of hyper-parameters is based on either 18 (bottom) or 10018 (top) players. Right: Parameter estimates for 18 real players when EB estimation of hyper-parameters is based on 10018 players, using as prior either a 3-component mixture of Gaussians (bottom) or a Gaussian (top).

Figure 2: X-axis: True event probability $q_j$; Y-axis: mean coverage of 95% posterior intervals for event probability. Mean is estimated by moving average. Case: $G = 2, p_g = 10, p = G \times p_G = 20, n_{\text{train}} = 100$. Methods: Hyb, EB, FB
Figure 3: X-axis: True event probability $q_j$; Y-axis: mean coverage of 95% posterior intervals for event probability. Mean is estimated by moving average. Case: $G = 2, p_g = 20, p = G * p_G = 40, n_{train} = 100$. Methods: Hyb, EB, FB.

Figure 4: X-axis: True event probability $q_j$; Y-axis: mean coverage of 95% posterior intervals for event probability. Mean is estimated by moving average. Case: $G = 2, p_g = 30, p = G * p_G = 60, n_{train} = 100$. Methods: Hyb, EB, FB.
Figure 5: X-axis: True event probability $q_j$; Y-axis: mean coverage of 95% posterior intervals for event probability. Mean is estimated by moving average. Case: $G = 2, p_g = 40, p = G \ast p_G = 80, n_{\text{train}} = 100$. Methods: Hyb, EB, FB

Figure 6: X-axis: True event probability $q_j$; Y-axis: mean coverage of 95% posterior intervals for event probability. Mean is estimated by moving average. Case: $G = 2, p_g = 50, p = G \ast p_G = 50, n_{\text{train}} = 100$. Methods: Hyb, EB, FB
Figure 7: X-axis: True event probability $q_j$; Y-axis: mean coverage of 95% posterior intervals for event probability. Mean is estimated by moving average. Case: $G = 5, p_g = 10, p = G \times p_G = 50, n_{\text{train}} = 200$. Methods: Hyb, EB

Figure 8: X-axis: True event probability $q_j$; Y-axis: mean coverage of 95% posterior intervals for event probability. Mean is estimated by moving average. Case: $G = 5, p_g = 20, p = G \times p_G = 100, n_{\text{train}} = 200$. Methods: Hyb, EB
Figure 9: X-axis: True event probability $q_j$; Y-axis: mean coverage of 95% posterior intervals for event probability. Mean is estimated by moving average. Case: $G = 5, p_g = 40, p = G \cdot p_G = 200, n_{\text{train}} = 200$. Methods: Hyb, EB
References

Bhattacharya, A. et al. (2015). Fast sampling with Gaussian scale-mixture priors in high-dimensional regression. arXiv:1506.04778 [stat]. ArXiv: 1506.04778. URL http://arxiv.org/abs/1506.04778

Chib, S. (1995). Marginal likelihood from the Gibbs output. J. Amer. Statist. Assoc., 90, 1313–1321.

Efron, B. and Morris, C. (1975). Data analysis using Stein’s estimator and its generalizations. J. Amer. Statist. Assoc., 70, 311–319.

Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. Biometrika, 12, 134–139.

Li, Q. and Lin, N. (2010). The Bayesian elastic net. Bayesian Analysis, 5, 151–170.

Press, S.J. (1982). Applied Multivariate Analysis, 2nd ed. Dover Publications, New York.

Van de Wiel, M.A. et al. (2012). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics, 14, 113–128.

Van Houwelingen, H.C. (2014). The role of empirical Bayes methodology as a leading principle in modern medical statistics. Biom. J., 56, 919–932.