

Understanding the asymptotic behaviour of empirical Bayes procedures

Harry van Zanten

Eindhoven University of Technology

Joint work with

Botond Szabó (TU Eindhoven)
Aad van der Vaart (VU Amsterdam)

BNP 2011, Veracruz

Outline

- Some simulations
- Theoretical questions
- Results
- Additional questions & concluding remarks

Some simulations

Signal in white noise model

Unknown function $f \in L^2[0, 1]$.

Model:

$$dX_t = f(t) dt + \frac{1}{\sqrt{n}} dW_t.$$

Observations:

$$(X_t : t \in [0, 1]).$$

Goal: **recover f** .

Simulated data

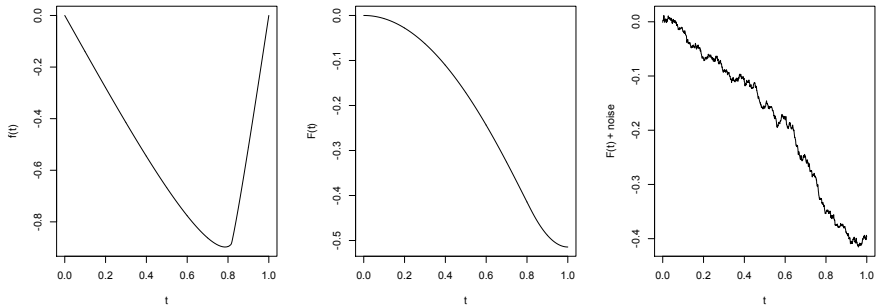


Figure: unknown function, its primitive and the noisy observations

Prior on f

Write $f = \sum f_k e_k$, for e_k an orthonormal basis of $L^2[0, 1]$. For instance

$$e_k(t) = \sqrt{2} \sin(k\pi t).$$

Put prior on Fourier coefficients f_k . We take

$$f_k \sim N(0, \tau^2 k^{-1-2\alpha}), \quad \text{independent.}$$

Here:

$\alpha > 0$ is fixed (“baseline” regularity of the prior)
scaling parameter τ is a **hyperparameter**.

Denote the resulting prior on f by Π_τ .

Empirical Bayes for the scaling parameter - 1

Let

$$X_k = \int_0^1 e_k(t) dX_t$$

be the observed noisy versions of the Fourier coefficients of f .

If

$$f \sim \Pi_\tau, \quad X | f \sim dX_t = f(t) dt + \frac{1}{\sqrt{n}} dW_t,$$

then

$$X_k \sim N\left(0, \frac{\tau^2}{k^{1+2\alpha}} + \frac{1}{n}\right)$$

and the X_k 's are independent.

Empirical Bayes for the scaling parameter - 2

Corresponding log-likelihood for τ :

$$\ell_n(\tau) = -\frac{1}{2} \sum_{k=1}^{\infty} \left(\log \left(1 + \frac{\tau^2 n}{k^{1+2\alpha}} \right) - \frac{\tau^2 n^2}{k^{1+2\alpha} + \tau^2 n} X_k^2 \right).$$

Empirical Bayes procedure:

1. Compute the posterior $\Pi_{\tau}(\cdot | X)$ corresponding to the prior Π_{τ} .
2. Substitute τ by the maximizer

$$\hat{\tau}_n = \operatorname{argmax}_{\tau > 0} \ell_n(\tau).$$

Estimate for f , empirical Bayes scaling

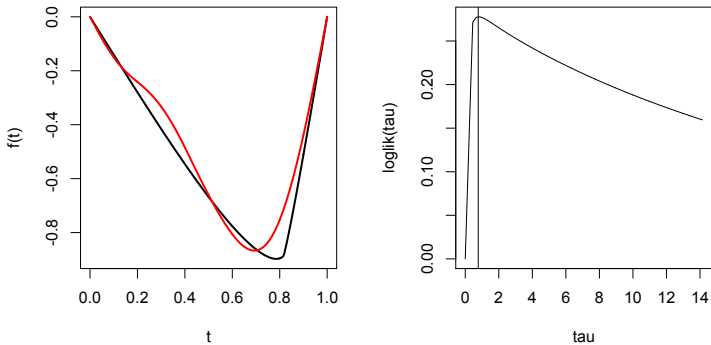


Figure: truth/posterior mean, likelihood for scaling parameter

Estimates for f , different scalings - 1

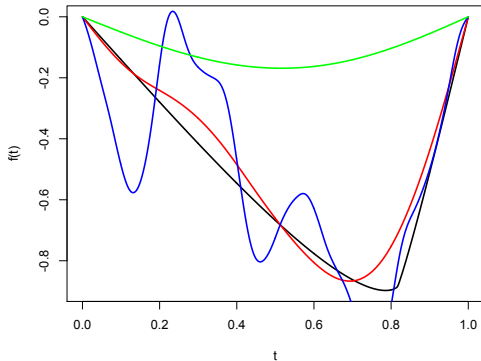


Figure: truth/posterior means

Estimates for f , different scalings - 2

Repeat 1000 times. Every time compute squared error $\|\hat{f} - f\|^2$ for each of the three estimators.

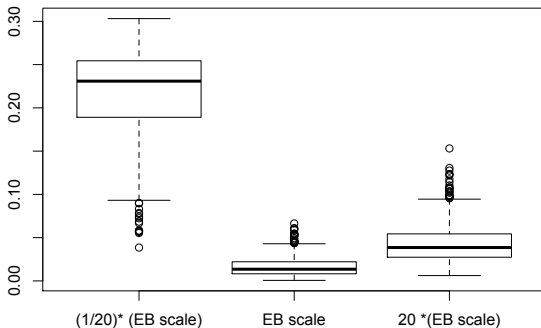


Figure: boxplot of the squared errors

Effect of different baseline regularities - 1

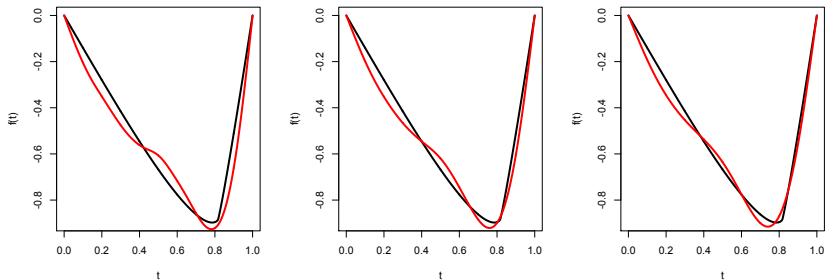


Figure: $\alpha = \beta, \beta + 1/2, \beta + 1$

Effect of different baseline regularities - 2

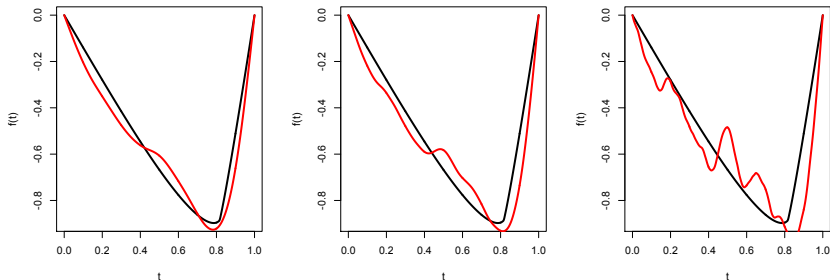


Figure: $\alpha = \beta, \beta - 1/2, \beta - 1$

Observations

1. Empirical Bayes seems to be able to do a good job at choosing the appropriate scaling.
2. Taking the initial prior too smooth does not seem to hurt too much.
3. Taking it too rough seems to deteriorate performance.

Can we make sense of this?

Can we prove theorems about this?

Observations

1. Empirical Bayes seems to be able to do a good job at choosing the appropriate scaling.
2. Taking the initial prior too smooth does not seem to hurt too much.
3. Taking it too rough seems to deteriorate performance.

Can we make sense of this?

Can we prove theorems about this?

Observations

1. Empirical Bayes seems to be able to do a good job at choosing the appropriate scaling.
2. Taking the initial prior too smooth does not seem to hurt too much.
3. Taking it too rough seems to deteriorate performance.

Can we make sense of this?

Can we prove theorems about this?

Observations

1. Empirical Bayes seems to be able to do a good job at choosing the appropriate scaling.
2. Taking the initial prior too smooth does not seem to hurt too much.
3. Taking it too rough seems to deteriorate performance.

Can we make sense of this?

Can we prove theorems about this?

Observations

1. Empirical Bayes seems to be able to do a good job at choosing the appropriate scaling.
2. Taking the initial prior too smooth does not seem to hurt too much.
3. Taking it too rough seems to deteriorate performance.

Can we make sense of this?

Can we prove theorems about this?

Observations

1. Empirical Bayes seems to be able to do a good job at choosing the appropriate scaling.
2. Taking the initial prior too smooth does not seem to hurt too much.
3. Taking it too rough seems to deteriorate performance.

Can we make sense of this?

Can we prove theorems about this?

Theoretical framework

Model and prior

Observations:

$$X_k = \theta_{0,k} + \frac{1}{\sqrt{n}} Z_k, \quad k = 1, 2, \dots$$

Here $\theta_0 \in \ell^2$ and Z_1, Z_2, \dots independent standard Gaussians.

Assume θ_0 in a **hyperrectangle**:

$$\theta_{0,k}^2 \leq C^2 k^{-1-2\beta}$$

for (unknown) $\beta, C > 0$.

Prior:

$$\Pi_\tau = \bigotimes_k N(0, \tau^2 k^{-1-2\alpha}).$$

Empirical Bayes posterior

Posterior:

$$\Pi_{\tau}(\cdot | X) = \bigotimes_k N\left(\frac{n}{n + k^{1+2\alpha}/\tau^2} X_k, \frac{\tau^2}{n\tau^2 + k^{1+2\alpha}}\right).$$

Log-likelihood for τ :

$$\ell_n(\tau) = -\frac{1}{2} \sum_{k=1}^{\infty} \left(\log\left(1 + \frac{\tau^2 n}{k^{1+2\alpha}}\right) - \frac{\tau^2 n^2}{k^{1+2\alpha} + \tau^2 n} X_k^2 \right).$$

EB posterior: Replace τ by maximizer $\hat{\tau}_n$ of ℓ_n :

$$\Pi_{\hat{\tau}_n}(B | X) = \Pi_{\tau}(B | X) \Big|_{\tau=\hat{\tau}_n}.$$

Oracle vs EB - 1

Recall:

- β : “regularity” of true sequence, α : “regularity” of prior.
- Minimax rate: $n^{-\beta/(1+2\beta)}$.
- For **fixed** scaling τ : optimal rate $n^{-\beta/(1+2\beta)}$ attained iff $\alpha = \beta$.

When allowing scaling $\tau = \tau_n$ such that $\tau_n \rightarrow 0$ or $\tau_n \rightarrow \infty$:

- Can attain optimal rate iff $\beta \leq 1 + 2\alpha$.
- Scaling $\tau_n = n^{(\alpha-\beta)/(1+2\beta)}$ yield optimal rate $n^{-\beta/(1+2\beta)}$.

Oracle vs EB - 1

Recall:

- β : “regularity” of true sequence, α : “regularity” of prior.
- Minimax rate: $n^{-\beta/(1+2\beta)}$.
- For fixed scaling τ : optimal rate $n^{-\beta/(1+2\beta)}$ attained iff $\alpha = \beta$.

When allowing scaling $\tau = \tau_n$ such that $\tau_n \rightarrow 0$ or $\tau_n \rightarrow \infty$:

- Can attain optimal rate iff $\beta \leq 1 + 2\alpha$.
- Scaling $\tau_n = n^{(\alpha-\beta)/(1+2\beta)}$ yield optimal rate $n^{-\beta/(1+2\beta)}$.

Oracle vs EB - 2

So in the range $\beta < 1 + 2\alpha$, a β -regular signal can be estimated optimally using the α -regular prior Π_{τ_n} , with τ_n the **oracle rescaling rate** $\tau_n = n^{(\alpha-\beta)/(1+2\beta)}$.

Main question: how does EB perform compared to the oracle?

Measure of performance: **contraction rate**. We say that the EB posterior **contracts around θ_0 at the rate ε_n** if

$$\Pi_{\hat{\tau}_n}(\theta : \|\theta - \theta_0\|_2 \leq M_n \varepsilon_n \mid X) \xrightarrow{\mathbb{P}_0} 1,$$

for every sequence $M_n \rightarrow \infty$.

Oracle vs EB - 2

So in the range $\beta < 1 + 2\alpha$, a β -regular signal can be estimated optimally using the α -regular prior Π_{τ_n} , with τ_n the **oracle rescaling rate** $\tau_n = n^{(\alpha-\beta)/(1+2\beta)}$.

Main question: how does EB perform compared to the oracle?

Measure of performance: **contraction rate**. We say that the EB posterior **contracts around θ_0 at the rate ε_n** if

$$\Pi_{\hat{\tau}_n}(\theta : \|\theta - \theta_0\|_2 \leq M_n \varepsilon_n \mid \mathbf{X}) \xrightarrow{\mathbb{P}_0} 1,$$

for every sequence $M_n \rightarrow \infty$.

Results

Asymptotics for the EB rescaling rate - 1

Asymptotic behaviour of $\hat{\tau}_n$ depends on θ_0 (like it should).

Define, for $\theta_0 \neq 0$,

$$h_n(\tau) = \sum_{k=1}^{\infty} \frac{(\tau^2 n)^{\frac{2\alpha}{1+2\alpha}} k^{1+2\alpha} \theta_{0,k}^2}{(k^{1+2\alpha} + \tau^2 n)^2},$$

and for $0 < l < L$,

$$\bar{\tau}_n = \sup \left\{ \tau > 0 : h_n(\tau) \geq l/n \right\},$$

$$\underline{\tau}_n = \sup \left\{ \tau > 0 : h_n(\tau) \geq L/n \right\}.$$

Asymptotics for the EB rescaling rate - 2

Asymptotically, $\hat{\tau}_n \in [\underline{\tau}_n, \bar{\tau}_n]$:

Theorem.

If $\theta_0 \neq 0$, then for small enough l and large enough L ,

$$\mathbb{P}_0(\underline{\tau}_n < \hat{\tau}_n < \bar{\tau}_n) \rightarrow 1.$$

(If $\theta_0 = 0$, then $\hat{\tau}_n = O_P(1/n)$.)

Asymptotics for the EB rescaling rate - 3

Proof.

Careful analysis of the marginal likelihood ℓ_n for τ .

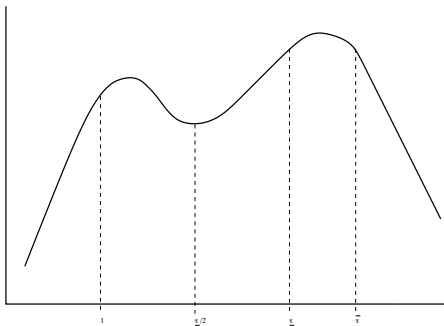


Figure: Shape of ℓ_n

Asymptotics for the EB rescaling rate - 4

Consider a “regular” signal: $\theta_{0,k}^2 \asymp k^{-1-2\beta}$.

$$\underline{\tau}_n \asymp \bar{\tau}_n \asymp \begin{cases} n^{\frac{\alpha-\beta}{1+2\beta}}, & \text{if } \beta < \alpha + 1/2, \\ n^{\frac{-1}{4+4\alpha}} (\log n)^{\alpha/(4+4\alpha)}, & \text{if } \beta = \alpha + 1/2, \\ n^{\frac{-1}{4+4\alpha}}, & \text{if } \beta > \alpha + 1/2. \end{cases}$$

So EB only selects the oracle rescaling rate if $\beta < \alpha + 1/2$!

In the other case, we expect sub-optimal performance.

Asymptotics for the EB rescaling rate - 4

Consider a “regular” signal: $\theta_{0,k}^2 \asymp k^{-1-2\beta}$.

$$\underline{\tau}_n \asymp \bar{\tau}_n \asymp \begin{cases} n^{\frac{\alpha-\beta}{1+2\beta}}, & \text{if } \beta < \alpha + 1/2, \\ n^{\frac{-1}{4+4\alpha}} (\log n)^{\alpha/(4+4\alpha)}, & \text{if } \beta = \alpha + 1/2, \\ n^{\frac{-1}{4+4\alpha}}, & \text{if } \beta > \alpha + 1/2. \end{cases}$$

So EB only selects the oracle rescaling rate if $\beta < \alpha + 1/2$!

In the other case, we expect sub-optimal performance.

Asymptotics for the EB posterior - 1

Theorem.

If $\theta_{0,k}^2 \leq C^2 k^{-1-2\beta}$ for all k , then

$$\Pi_{\hat{\tau}_n}(\theta : \|\theta - \theta_0\|_2 \leq M_n \varepsilon_n | \mathbf{X}) \xrightarrow{\mathbb{P}_0} 1,$$

for every sequence $M_n \rightarrow \infty$, where

$$\varepsilon_n = \begin{cases} n^{-\beta/(1+2\beta)}, & \text{if } \beta < 1/2 + \alpha, \\ n^{-\beta/(1+2\beta)} (\log n)^p, & \text{if } \beta = 1/2 + \alpha, \\ n^{-(1/2+\alpha)/(2+2\alpha)}, & \text{if } \beta > 1/2 + \alpha. \end{cases}$$

If also $\theta_{0,k}^2 \geq c^2 k^{-1-2\beta}$ for $c > 0$, then, for all sufficiently small m ,

$$\Pi_{\hat{\tau}_n}(\theta : \|\theta - \theta_0\|_2 < m \varepsilon_n | \mathbf{X}) \xrightarrow{\mathbb{P}_0} 1.$$

Asymptotics for the EB posterior - 2

Proof.

- We know that $\hat{\tau}_n \in [\underline{\tau}_n, \overline{\tau}_n]$ w.p. to 1.
- We prove that

$$\mathbb{E}_0 \sup_{\tau \in [\underline{\tau}_n, \overline{\tau}_n]} \int \|\theta - \theta_0\|_2^2 \Pi(d\theta | \mathbf{X}) \lesssim \varepsilon_n^2.$$

Asymptotics for the EB posterior - 3

Observations:

- If $\beta < \alpha + 1/2$, the EB procedure **matches the oracle**.
- If $\alpha + 1/2 \leq \beta \leq 2\alpha + 1$, EB performs **worse than the oracle**.
- If $2\alpha + 1 < \beta$, even the oracle performs sub-optimally.
- Results are sharp.

There is a cut-off at $\alpha + 1/2$ (= “RKHS regularity” of the prior...)

Additional questions & concluding remarks

Main messages so far

- Empirical Bayes methods can yield **adaptive, rate-optimal** procedures.
- For good performance, the unscaled prior should be **sufficiently regular** relative to the truth.
- Too much undersmoothing yields **sub-optimal rates**.

Some first answers - 1

Can we get good results for **every** $\beta > 0$ by rescaling an “infinitely smooth” prior?

Attempt 1: use prior

$$\Pi_\tau = \bigotimes_k N(0, \tau^2 e^{-k})$$

and select τ by EB.

There exists oracle scaling that yields optimal rates for all $\beta > 0$.

EB always oversmooths: **suboptimal rates!**

Some first answers - 1

Can we get good results for **every** $\beta > 0$ by rescaling an “infinitely smooth” prior?

Attempt 1: use prior

$$\Pi_\tau = \bigotimes_k N(0, \tau^2 e^{-k})$$

and select τ by EB.

There exists oracle scaling that yields optimal rates for all $\beta > 0$.

EB always oversmooths: **suboptimal rates!**

Some first answers - 1

Can we get good results for **every** $\beta > 0$ by rescaling an “infinitely smooth” prior?

Attempt 1: use prior

$$\Pi_\tau = \bigotimes_k N(0, \tau^2 e^{-k})$$

and select τ by EB.

There exists oracle scaling that yields optimal rates for all $\beta > 0$.

EB always oversmooths: **suboptimal rates!**

Some first answers - 1

Can we get good results for **every** $\beta > 0$ by rescaling an “infinitely smooth” prior?

Attempt 1: use prior

$$\Pi_\tau = \bigotimes_k N(0, \tau^2 e^{-k})$$

and select τ by EB.

There exists oracle scaling that yields optimal rates for all $\beta > 0$.

EB always oversmooths: **suboptimal rates!**

Some first answers - 2

Attempt 2: use prior

$$\Pi_\tau = \bigotimes_k N(0, e^{-\tau k})$$

and select τ by EB.

This works! EB achieves optimal rates for all $\beta > 0$.

So yes, we can adapt to all smoothness levels $\beta > 0$ with a single prior. But care is needed in the construction of the prior.

Some first answers - 2

Attempt 2: use prior

$$\Pi_\tau = \bigotimes_k N(0, e^{-\tau k})$$

and select τ by EB.

This works! EB achieves **optimal rates** for **all** $\beta > 0$.

So **yes**, we can adapt to all smoothness levels $\beta > 0$ with a single prior. But care is needed in the construction of the prior.

Some first answers - 2

Attempt 2: use prior

$$\Pi_\tau = \bigotimes_k N(0, e^{-\tau k})$$

and select τ by EB.

This works! EB achieves **optimal rates** for **all** $\beta > 0$.

So **yes**, we can adapt to all smoothness levels $\beta > 0$ with a single prior. But care is needed in the construction of the prior.

Main messages

- Empirical Bayes methods can yield **adaptive, rate-optimal** procedures.
- For good performance, the unscaled prior should be **sufficiently regular** relative to the truth.
- Too much undersmoothing yields **sub-optimal rates**.
- Many aspects still unclear.

TO BE CONTINUED...