Statistica Sinica **22** (2012), 1217-1232
doi:http://dx.doi.org/10.5705/ss.2010.014

HIGH DIMENSIONAL EXPONENTIAL FAMILY ESTIMATION VIA EMPIRICAL BAYES

Omkar Muralidharan

Stanford University

Abstract: Consider estimating $\theta_1, \ldots, \theta_N$ based on data $z_i \sim f_{\theta_i}$, where f_θ is a continuous natural exponential family. One successful approach is Empirical Bayes (EB). EB methods assume that θ come from an unknown prior, and estimate the Bayes procedure corresponding to that prior.

In this paper, we propose a general form of nonparametric EB estimator that uses estimates of the marginal density of z and its derivative. This estimator was first proposed by Zhang (1997) for the normal means problem, $z_i \sim \mathcal{N}(\theta_i, 1)$. We bound the regret of our method in terms of the error in estimating the marginal density and its derivative. As a side point, our proof yields a lower bound on the regret of general estimators.

We illustrate our method in the simultaneous chi-squared problem, where z_i is a chi-squared random variable with scale $1/\theta_i$. We specialize our theoretical results to this case, and study the empirical performance of our method under different estimators of the marginal and its derivative. Our method outperforms the UMVU estimator and a conjugate prior parametric EB approach.

Key words and phrases: Chi-squared estimation, density estimation, empirical Bayes, exponential family, regret bound, separable estimator.

1. Introduction

Suppose we have independent, real-valued, data $z_1, \ldots, z_N$, and each z_i has distribution f_{θ_i}, where

$$f_\theta(z) = \exp\left(\theta z - \psi(\theta)\right) f_0(z)$$

is an absolutely continuous natural exponential family. We want to estimate $\theta_1, \ldots, \theta_N$ under squared error loss. It is well known that the MLE is a poor estimator of θ when N is large.

Empirical Bayes (EB) methods have been successfully used to construct better estimators. EB methods model θ_i as iid from a prior G, making the full model

$$\begin{aligned} \theta_i &\sim G \\ z_i|\theta_i &\sim f_{\theta_i} \end{aligned}$$

independently for $i = 1, \ldots, N$. In this model, we want an estimator $\hat{\theta} = t(z)$ that minimizes the Bayes risk $\mathcal{R}(t, G) = E_G\left((t(z) - \theta)^2\right)$. For a fixed prior G, Robbins (1954) showed that the Bayes estimator is

$$t_G(z) \equiv E_G(\theta|z) = -\frac{f_0'(z)}{f_0(z)} + \frac{f_G'(z)}{f_G(z)}, \tag{1.1}$$

where f_G is the marginal distribution of z and f_0 is the carrier density of the family. This result holds as long as f_0 is absolutely continuous Berger (1980).

EB methods treat G as unknown and estimate t_G. Parametric EB methods assume G (or t_G) has a certain parametric form, then fit it using the data. Nonparametric EB (NPEB) methods attempt to estimate t_G consistently for all G. The EB approach has produced estimators with good theoretical and practical properties (Singh (1979); Singh and Wei (1992); Brown and Greenshtein (2009); Jiang and Zhang (2009); Zhang (1997, 2003)).

In this paper, we propose a general form of NPEB estimator. Based on estimates $\hat{f}$ and $\hat{f}'$ of f_G and f_G', we suggest using the estimator

$$\hat{t}_\rho = -\frac{f_0'}{f_0} + \frac{\hat{f}'}{\hat{f} \vee \rho}.$$

This estimator is simple: we just plug $\hat{f}$ and $\hat{f}'$ into Robbins' formula, but to avoid dividing by a near-zero quantity, we replace $\hat{f}$ by $\hat{f} \vee \rho = \max(\hat{f}, \rho)$. This kind of estimate was suggested for the normal problem ($z \sim \mathcal{N}(\theta, 1)$) by Zhang (1997) and studied further by Jiang and Zhang (2009).

Their methods, however, are readily generalized to all exponential families with absolutely continuous f_0, that is, to all families where Robbins' formula holds. We bound the regret of $\hat{t}_\rho$ (the difference in risk between $\hat{t}_\rho$ and t_G) in terms of the error in $\hat{f}$ and $\hat{f}'$. The bounds can then be used to show that $\hat{t}_\rho$ asymptotically achieves the Bayes risk. Our proof technique also yields a lower bound for the regret of a general estimator in terms of a related density estimate. This lower bound further motivates NPEB methods and provides a useful diagnostic for general methods.

We illustrate the general theory by applying it to the simultaneous chi-squared estimation problem, where z_i comes from a chi-squared distribution with inverse scale θ_i. This problem arises in microarray data: the θ_i are needed to construct t-statistics, which in turn are used in many simultaneous inference procedures to find differentially expressed genes Efron (2008). We study the finite-sample performance of $\hat{t}_\rho$ for different choices of $\hat{f}$ and $\hat{f}'$ by simulation, and specialize our theoretical results to the chi-squared case.

2. Proposed Method and Regret Bounds

In this section, we outline our proposed NPEB method and prove bounds on the regret. We also obtain lower bounds on the regret of a general estimator for this problem; this side point motivates NPEB methods and provides a useful diagnostic.

2.1. Setup, regret and the proposed estimator

For the rest of the paper, we work with the Bayesian model

$$\theta \sim G$$
$$z|\theta \sim f_\theta.$$

We assume that there are N previous draws (θ_i, z_i) from this model, and we observe $z_1, \ldots, z_N$. We want to estimate θ based on z for a new draw (θ, z). We use $z_1, \ldots, z_N$ to construct an estimator $t(z) = t(z; z_1, \ldots, z_N)$. We condition on $z_1, \ldots, z_N$ throughout; all expectations and probabilities are conditional on $z_1, \ldots, z_N$, though our notation supresses this. This lets us treat $t(z)$ as a fixed function of z.

Our goal is to construct an estimator that performs nearly as well as the Bayes estimator t_G. Let the Bayes risk of an estimator $t(z)$ be

$$\mathcal{R}(t, G) = E_G\left((\theta - t(z))^2\right).$$

The *regret* of t is how much extra Bayes risk we get by using t instead of the Bayes estimator t_G:

$$\Delta(t, G) = \mathcal{R}(t, G) - \mathcal{R}(t_G, G).$$

Because we use squared-error loss, t_G is just $E_G(\theta|z)$, and the definition of conditional expectation implies that

$$\Delta(t, G) = E_G\left((t(z) - t_G(z))^2\right). \tag{2.1}$$

This actually holds even if θ is not square integrable, as long as $\mathcal{R}(t_G, G)$ is finite Singh (1979); Brown (1971). Acheiving low regret is thus equivalent to estimating t_G well under squared error loss.

We propose estimating t_G using a tempered NPEB estimator. The most obvious approach based on equation (1.1) would be to use $z_1, \ldots, z_N$ to estimate f_G and f_G' by, say, $\hat{f}$ and $\hat{f}'$, then plug in to estimate t_G. But if $\hat{f}$ is too small, $\frac{\hat{f}'}{\hat{f}}$ may be too large, and we may overshrink. Zhang (1997) introduced a simple solution in the normal case - replace the $\hat{f}$ in our plug-in estimator by

$\hat{f} \vee \rho = \max\left(\hat{f}, \rho\right)$ for some small ρ, but keep $\hat{f}'$ the same. Using this approach for other exponential families gives us a tempered EB estimator

$$\hat{t}_\rho = -\frac{f_0'}{f_0} + \frac{\hat{f}'}{\hat{f} \vee \rho}. \tag{2.2}$$

Tempering protects us from overshrinking. In the tails, $\hat{f}' \to 0$ and $\hat{f} \vee \rho \to \rho$, so $\hat{f}'/(\hat{f} \vee \rho) \to 0$. So in the tails, $\hat{t}_\rho$ approaches $-f_0'/f_0$, the UMVU estimator of θ Sharma (1973). This is sensible, since the tails are exactly where we have the least information about f_G. Tempered EB estimators are similar to the limited translation estimators introduced by Efron and Morris (1971).

2.2. Bounding the regret

We now bound the regret of $\hat{t}_\rho$ in terms of the error in $\hat{f}$ and $\hat{f}'$. Our bounds generalize results of Zhang (1997) and Jiang and Zhang (2009). Recall that we are conditioning on $z_1, \ldots, z_N$, so that regret is a conditional expectation and $\hat{f}$ and $\hat{f}'$ are fixed functions of z.

Lemma 1. *Suppose that* $f_G'/(f_G \vee \rho) \le A(\rho)$. *Then*

$$\Delta\left(\hat{t}_\rho, G\right)^{1/2} \le \frac{1}{\rho}\left(\int \left(\hat{f}' - f_G'\right)^2 f_G dz\right)^{1/2} + \frac{1}{\rho} A(\rho)\left(\int \left(\hat{f} - f_G\right)^2 f_G dz\right)^{1/2} + T(\rho, f_G),$$

where $T(\rho, g) = (\int (1 - g/\rho)_+^2 (g'/g)^2 f_G dz)^{1/2}$.

Lemma 1 has two unfamiliar features, a tempering term and a bound $A(\rho)$. The tempering term $T(\rho, f_G)$ depends on the heaviness of the tail of f_G and behaves roughly like $\rho^{1/2}$. If f_G has exponential or lighter tails, it behaves like $\rho^{1/2}$ with some log factors, and if f_G falls as z^{-k}, it behaves like $\rho^{1/2-1/2k}$. The bound $A(\rho)$ measures how quickly f_G' drops off compared to f_G. We always have $A(\rho) \le (1/\rho) \sup \|f_\theta'\|_\infty$, but sometimes we can do better. In the normal case, Jiang and Zhang (2009) get $A(\rho) = \mathcal{O}(\log \rho)$.

Lemma 1 bounds the regret by error in $\hat{f}$ and $\hat{f}'$. If $\hat{f}$ and f_G are smooth, we can reduce this to a bound in terms of the error in $\hat{f}$, since if $\hat{f}$ and f_G are smooth and $\hat{f}$ is close to f_G, $\hat{f}'$ should be close to f_G'. Theorem 1, below, makes this precise.

The right kind of smoothness turns out to be the decay of the Fourier transforms of the densities. Sometimes z is not supported on the whole real line, so it is natural for f_G and $\hat{f}$ to have discontinuities at the boundary of the support, giving their Fourier transforms heavy tails. In this case, the theorem can be

applied to smooth extensions of f_G and $\hat{f}$ that agree with the originals on the support of z. Theorem 1 is conditional on $z_1, \ldots, z_N$, but only requires that our particular realization of $\hat{f}$ be smooth.

Theorem 1. *Let f^* be the Fourier transform of a function f. Suppose $|f_G^*(u)|$, $|\hat{f}^*(u)| \leq H(u)$ for almost all $|u| \geq C$, where $\int u^2 H(u)^2 < \infty$. Let $L(a) = (1/a^2)\int_{|u|\geq a} u^2 H(u)^2\, du$; $L(a) \downarrow 0$ as $a \to \infty$. Then*

$$\Delta(\hat{t}_\rho, G)^{1/2} \leq \frac{1}{\rho}\|f_G\|_\infty^{1/2}\left(\sqrt{\frac{5}{2\pi}}L^{-1}\left(d(\hat{f}, f_G)^2\right) + A(\rho)\right)d(\hat{f}, f_G) + T(\rho, f_G),$$

where $d(\hat{f}, f_G) = (\int(\hat{f} - f_G)^2 dz)^{1/2}$.

Theorem 1 generalizes a result of Jiang and Zhang (2009) from the normal case to exponential families and more general density estimators. It shows that if our densities are smooth, the regret is bounded by the density estimation error, up to smoothness and tempering terms. The tempering term is the same as in Lemma 1. The smoothness term $L^{-1}(d^2)$ depends on how fast the characteristic functions of f_G and $\hat{f}$ decay: if exponentially, $L^{-1}(d^2)$ behaves like $\log d$; if they decay as u^{-k}, it behaves like $d^{-2/(2k-1)}$. Since we are conditioning on $z_1, \ldots, z_N$, what matters is the smoothness of f_G and the realized $\hat{f}$. Theorem 1 is not as sharp as the result of Jiang and Zhang (2009) when applied to the normal case.

2.3. Aside: A lower bound on the regret

Our proofs lead to another motivation for NPEB methods and a useful diagnostic tool for the simultaneous estimation problem. Consider a general estimator

$$t(z) = -\frac{f_0'}{f_0} + \frac{f_t'}{f_t},$$

where $\log f_t = \int_0^z (t(x) + f_0'(x)/f_0(x))dx$. We can view t as coming from Robbins' formula with f_t plugged in as an estimate of f_G; roughly speaking, f_t is the marginal density of z that would make t Bayes, though f_t may not be a proper density. Working backward from an estimator to a marginal density using Robbins' formula has previously been used to prove admissibility results in the normal case (Brown and Greenshtein (2009); Berger and Srinivasan (1978)).

Theorem 2 shows that for t to have low regret f_t must be close to f_G. Since f_t is only determined up to scale, we fix $f_t(z_0) = f_G(z_0)$ at some arbitrary point z_0.

Theorem 2. *Let F_G be the cdf of z under G and*

$$P(z) = \begin{cases} \frac{F_G(z)}{f_G(z)} & z \leq z_0, \\ \frac{1-F_G(z)}{f_G(z)} & z \geq z_0. \end{cases}$$

Then if f_t is scaled so $f_t(z_0) = f_G(z_0)$,

$$\int \left|\log \frac{f_G}{f_t}\right| f_G dz \leq \left(\int P(z)^2 f_G(z) dz\right)^{1/2} \Delta(t, G)^{1/2}.$$

Alternatively, if f_t is integrable and scaled to be a density, a simple modification of Theorem 2 shows that

$$D_{KL}(f_t \| f_G) \leq \inf_{z_0} \left[\left(\int P(z)^2 f_G(z) dz\right)^{1/2} \Delta(t, G)^{1/2} + \log \frac{f_t(z_0)}{f_G(z_0)}\right],$$

where $D_{KL}(f\|g) = \int \log \frac{g}{f} g dz$ is the Kullback-Liebler divergence. Versions of Theorem 2 also hold for the tempered NPEB method that is the main focus of this paper.

Theorem 2 suggests that if we restrict our attention to EB methods based on estimates of f_G, we do not overlook different techniques with low regret. It also provides a diagnostic. Given an estimator t, we can work out f_t, and see if it matches the observed distribution of the data; a glaring mismatch indicates that t has high regret. For example, consider soft-thresholding in the normal case ($f_\theta = \mathcal{N}(\theta, 1)$). If $t(z) = \text{sign}(z)(|z| - \lambda)_+$, f_t is a Huber density that transitions from normal to exponential at the threshold λ. If $z_1, \ldots, z_N$ look unlikely to have come from a Huber density, t may well be outperformed by methods that match f_G more closely.

3. Application: Estimating Inverse Variance

3.1. Specializing theoretical results

We illustrate our method in the simultaneous chi-squared problem, where z_i is a chi-squared random variable with k degrees of freedom and scale $1/\theta_i$. The distribution of $z|\theta$ is

$$f_\theta(z) = C_k \theta^{n/2} z^{n/2-1} \exp\left(-\frac{\theta z}{2}\right).$$

This is an exponential family with natural parameter θ and sufficient statistic $-z/2$. We try to estimate θ well under squared error loss. There are other loss functions that may be of more interest, for example, the loss functions considered by Berger (1980). Extending our results to more general losses seems difficult, as we discuss in the conclusion.

The previous theory is easily adapted to give estimates of θ based on z. Robbins' formula is

$$E(\theta|z) = \frac{k-2}{z} - 2\frac{f_G'(z)}{f_G(z)},$$

where the factor of -2 comes from the fact that the sufficient statistic is $-z/2$, not z. We construct $\hat{t}_\rho$ by constructing estimates $\hat{f}$ and $\hat{f}'$, then plugging in to get $\hat{t}_\rho = (k-2)/z - 2\hat{f}'/(\hat{f} \vee \rho)$. Corollary 1 specializes Theorem 1 to the chi-squared problem.

Corollary 1. *Suppose G is integrable and $k \geq 5$. Suppose that for some $\alpha \in (0, 1-4/k)$, $P_G(\theta \geq m) \leq Em^{-[(1-\alpha)/\alpha][k/2]}$ for all $m \geq M$, and $|\hat{f}^*(u)| \leq Bu^{-(1-\alpha)(k/2)}$ for $u \geq C$. Then for some constant F that depends on $E_G(\theta)$, B, C and M,*

$$\Delta\left(\hat{t}_\rho, G\right)^{1/2} \leq \frac{F}{\rho}\left(d\left(\hat{f}, f_G\right)^{-2/(1+(1-\alpha)k)} + A(\rho)\right)\left(d\left(\hat{f}, f_G\right)\right) + T(\rho, f_G).$$

The condition $P_G(\theta \geq m) = \mathcal{O}(m^{-[(1-\alpha)/\alpha](k/2)})$ is satisfied, for example, if G has tails like a Gamma distribution. In this case, the smoothness of $\hat{f}$ becomes the limiting factor; requiring that $\hat{f}^*$ decays as $u^{-\gamma}$ roughly corresponds to $\hat{f}$ having continuous γth derivative. Interestingly, the constant in Corollary 1 only depends on G through its mean and tail behavior. The corollary thus holds uniformly in classes of priors with bounded mean and constrained tail behavior.

If G is bounded and bounded away from 0, and $\hat{f}$ is smooth enough, we can give the rate at which the regret converges more explicitly. The constants in Corollary 2 only depend on the support of G, so the result holds uniformly across all priors with the same support.

Corollary 2. *Suppose $k \geq 5$, $0 < M_1 \leq \theta \leq M_2 < \infty$, $|\hat{f}^*(u)| \leq Bu^{-k/2}$ for all $u \geq C$. Then if we choose $\rho = \mathcal{O}(d(\hat{f}, f_G)^{2/5})$, $\Delta(\hat{t}_\rho, G) = \mathcal{O}(d(\hat{f}, f_G)^{2/5})$, with constants that depend on M_1, M_2, B and C.*

3.2. An empirical comparison

We compared our tempered NPEB estimator to the UMVU estimator, a conjugate-prior parametric EB estimator, and an estimator introduced by Berger (1980). Our theory used a sequential setup wherein we observed $z_1, \ldots, z_N$ and estimated θ for a new observation z. Our simulations use the more realistic situation where we observe $z_1, \ldots, z_N$ and estimate $\theta_1, \ldots, \theta_N$. Since each z_i can be treated as the "new observation," and each z_i only affects the density estimate slightly, our theory still applies.

3.2.1. The UMVU estimator and Berger's estimator

The UMVU estimator for θ is $(k-2)/z$; it is the multiple of $1/z$ with lowest mean-squared-error, dominating the MLE k/z (the MLE is also the Jeffrey's prior

posterior mean). Berger (1980) found an estimator that dominates the UMVU estimator:

$$\hat{\theta}_i = \frac{k-2}{z_i} + \frac{cz_i}{b + \sum z_i^2},$$

where $b \geq 0$ and $c \in (0, 4\,(N-1))$. Berger left the choice of b and c open, so we tried many different values. All of them performed nearly the same, and none substantially improved on the UMVU estimator. We used $b = c = N$.

3.2.2. A parametric EB estimator

Our parametric EB estimator used a conjugate prior whose parameters were estimated by the method of moments. Efron and Morris (1973) take this approach in the normal case to obtain an empirical Bayes construction of the James-Stein estimator.

The conjugate prior for the chi-squared distribution is the Gamma distribution, $Gamma\,(\alpha, \beta)\,(x) = (1/\Gamma)\,(\alpha)\beta^{\alpha} x^{\alpha-1} \exp\,(-x\beta)$. If $\theta \sim Gamma\,(\alpha, \beta)$ and $z = (1/\theta)\chi_k^2$, then it is easy to show for $\alpha > 2$, that $E\,(z) = k[\beta/(\alpha-1)]$, $E\left(z^2\right) = k\,(k+2)\,(\beta^2/(\alpha-1)\,(\alpha-2))$, and $E\,(\theta|z) = ((z/2)/(\beta + z/2))\,(k/z) + (\beta/(\beta + z/2))\,(\alpha/\beta)$.

We estimated α, β by method of moments, then plugged in to estimate $E\,(\theta|z)$. With $m_1 = \bar{z}/k$ and $m_2 = \bar{z^2}/k\,(k+2)$, the method of moment estimates are $\hat{\alpha} = \max(1 + m_2/m_2 - m_1^2, 3)$ (we fixed$\hat{\alpha} \geq 3$ to ensure that z has finite variance) and $\hat{\beta} = m_1 m_2/(m_2 - m_1^2)$. Plugging these in gives an estimate of $E\,(\theta|z)$. If the prior is very concentrated, $m_2 - m_1^2$ can be negative. In this case we fit a very concentrated Gamma by taking $\hat{\alpha}$ to be essentially infinite (10^8) and $\hat{\beta} = \hat{\alpha}\bar{z}$.

3.2.3. Tempered NPEB estimators

Specifying the tempered NPEB estimator requires fixing $\hat{f}, \hat{f}'$ and ρ. We used two choices for $\hat{f}$ and $\hat{f}'$. The first estimator was an off-the-shelf log-spline estimator. The density estimate takes the form

$$\hat{f}\,(z) \propto \exp\left(\sum \beta_i c_i\,(z)\right),$$

where the $c_i\,(z)$ are a natural spline basis. We used the default natural spline basis supplied by R, with 15 degrees of freedom and boundary knots at the 1st and 99th percentiles of z. We fit $\hat{f}$ by binning z and fitting a Poisson GLM. Efron (2009) used this approach in the normal case; the reference contains details on the fitting method. Since cubic spines are smooth, the characteristic function of $\hat{f}$ should decay quickly. The cubic splines' discontinuous third derivative means that Corollary 2 does not apply, but the log-spline estimator seemed to perform well anyway.

We chose the degrees of freedom to give $\hat{f}$ enough flexibility to model all the test scenarios. We made the choice by plotting histograms of z and assessing the fit by eye, mimicking the process we would use with data.

Our second density estimator was a Gamma mixture model. We modeled the prior G as a mixture, $G = \sum \pi_i Gamma\,(a_i, b_i)$. We fixed a and b to a grid of values, then fit π by the EM algorithm. Details on the choice of a, b, and the EM algorithm are in the appendix. We used 10 mixture groups. As for the log-spline, we assessed the fit of $\hat{f}$ by eye, and chose the number of groups to to give $\hat{f}$ enough flexibility to model the test scenarios. Given the fitted $\hat{G}$, we used the density estimator $\hat{f} = f_{\hat{G}}$. Since $\hat{f}$ corresponds to a prior $\hat{G}$ with a Gamma-like tail, it is quite smooth, and both corollaries apply.

Both methods were insensitive to choice of ρ, as long as it was small. We took $\rho = 10^{-6}$, but $\rho = 0$ performed just as well.

3.2.4. Testing scenarios

We tested the methods under distributions of θ ranging from smooth to sparse. We used the following priors, shown in Figure 1.

1. $Gamma\,(10, 1)$, a smooth prior.
2. $Unif\,(2, 4)$, another smooth prior.
3. 50% $Gamma\,(10, 7)$, 50% $Gamma\,(10, 20)$, a smooth but bimodal prior.
4. 25% at $\theta = 1$, 50% at $\theta = 2$, and 25% at $\theta = 10$, a three point prior with one extreme point.
5. 75% $Gamma\,(1,000, 1,000)$, 25% $Gamma\,(1,000, 333)$, an approximate two-point prior
6. A point mass at $\theta = 1$.

We tested the methods with $k = 10$; increasing k improved all methods' performance, but did not substantially change their relative performance. We took $N = 10,000$, a size typical of microarray studies.

These priors are fair, in the sense that none of the methods are fitting the true model (except the parametric EB method on prior 1). The UMVU, Berger, parametric EB and log-spline estimators are clearly not tailored to these priors. The Gamma mixture method used a fixed grid of 10 gamma groups for a non-parametric fit; it does not have an unfair advantage in fitting the Gamma priors used here.

3.2.5. Results

Our simulation results are in Tables 1 and 2. Berger's estimator dominates the UMVU estimator, but its advantage is small. The parametric EB method does well on the smooth priors, including the smooth bimodal prior, but does

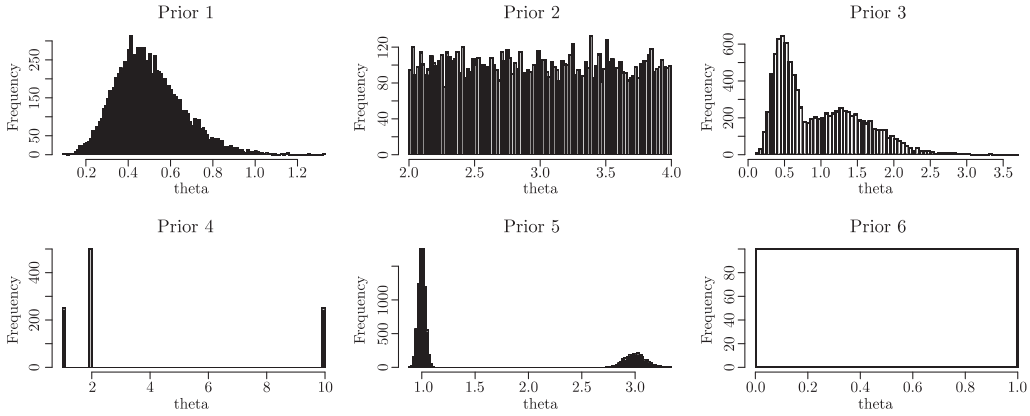

Figure 1. Simulation priors as described in the text.

Table 1. Mean squared errors $(1/N)\sum(\hat{\theta}-\theta)^2$ from the simulations under the priors in the text. The quantities shown are averages over 100 simulations, with standard deviations given in parentheses.

Method	Prior 1	Prior 2	Prior 3
UMVU	12.698 (0.4745)	1.167 (0.0351)	0.1577 (0.0070)
Berger	13.252 (0.4778)	1.114 (0.0352)	0.1556 (0.0070)
Parametric EB	5.239 (0.1002)	0.2417 (0.0031)	0.1002 (0.0020)
Log-spline NPEB	6.378 (0.2321)	0.3296 (0.0187)	0.1146 (0.0050)
Mixture NPEB	5.248 (0.1021)	0.0031 (0.0032)	0.0831 (0.0018)
Bayes	5.235 (0.1004)	0.0031 (0.0031)	0.0798 (0.0017)

Method	Prior 4	Prior 5	Prior 6
UMVU	3.3966 (0.1677)	0.3756 (0.0157)	0.125 (0.0034)
Berger	3.3786 (0.1678)	0.3668 (0.0157)	0.118 (0.0035)
Parametric EB	6.4130 (0.0479)	0.3970 (0.0045)	2.10×10^{-5} (2.9×10^{-5})
Log-spline NPEB	2.9670 (0.1396)	0.1830 (0.0099)	9.24×10^{-3} (2.8×10^{-3})
Mixture NPEB	0.9825 (0.0392)	0.1688 (0.0057)	3.21×10^{-5} (5.3×10^{-5})
Bayes	0.3218 (0.0354)	0.1241 (0.0051)	0 (0)

very badly on the point priors, doing even worse than the UMVU estimator. It does well, however, on the point mass, because the moment-based fitting can detect that the prior is extremely concentrated.

The NPEB methods did well on all the priors, and the mixture model did best. Both the mixture model and the parametric EB method essentially acheived the Bayes error on priors $1, 2$ and 6 (the smooth unimodal priors and the point mass), and the mixture model was much better on the rest. The log-spline estimator was not as good, but it did remarkably well for an off-the-shelf estimator. It was a bit worse than the parametric EB method and mixture model on priors

Table 2. Relative regret from the simulations under the priors in the text. The relative regret is $MSE(\hat{\theta})/MSE(\hat{\theta}_{bayes}) - 1$. The quantities shown are averages and standard deviations (in parentheses) over 100 simulations. The relative regret is infinite for all methods on prior 6, as the Bayes risk is 0.

Method	Prior 1	Prior 2	Prior 3
UMVU	1.6165 (0.083)	3.8593 (0.149)	0.9765 (0.085)
Berger	1.5313 (0.083)	3.6354 (0.149)	0.9490 (0.085)
Parametric EB	0.0007 (0.0001)	0.0059 (0.002)	0.2554 (0.021)
Log-spline NPEB	0.2181 (0.038)	0.3295 (0.075)	0.4362 (0.058)
Mixture Gamma NPEB	0.0023 (0.0018)	0.0114 (0.004)	0.0419 (0.009)

Method	Prior 4	Prior 5
UMVU	9.673 (1.234)	2.0290 (0.170)
Berger	9.627 (1.228)	1.9590 (0.169)
Parametric EB	19.167 (2.262)	2.2030 (0.125)
Log-spline NPEB	8.321 (1.061)	0.4753 (0.073)
Mixture Gamma NPEB	2.080 (0.273)	0.3584 (0.037)

$1, 2, 6$. On the rest, it trailed the mixture model but outperformed the others.

These results suggest that our tempered NPEB method can match a conjugate-prior parametric EB approach on smooth unimodal priors, and substantially outperform it when the prior is bimodal or sparse. Using the off-the-shelf log-spline yields good performance, but we can improve by using an appropriate density estimator for the problem, in this case the mixture model.

4. Conclusion

In this paper we proposed a tempered NPEB method based on estimates $\hat{f}$, $\hat{f}'$ of the marginal density f_G and its derivative f'_G. We proved that our method performs well if $\hat{f}$ and $\hat{f}'$ are good estimates. We illustrated our method on the simultaneous chi-squared estimation problem and our method performed well empirically.

Many questions remain. First, we considered only continuous exponential families. Robbins (1954) was largely concerned with discrete families, such as the Poisson, but his formula fails for discrete families. In the cases he considered, however, the Bayes estimator was still expressible in terms of the marginal distribution of the data. Our approach may extend to these cases, even if the present techniques do not.

Second, our method only applies to squared-error loss. In the chi-squared problem, for example, Berger (1980) suggested scaled loss functions of the form $\theta^m(1 - \hat{\theta}\theta)^2$. Our results depend heavily on Robbins' formula to express the posterior mean in terms of f_G and f'_G, but Bayes estimators for other loss functions are not so easily expressed. On the other hand, higher order versions of Robbins'

formula give the posterior cumulants of θ in terms of f_G and its derivatives. Our results may extend to losses for which the Bayes estimator is approximately a function of the first few posterior cumulants.

Acknowledgements

The author thanks Professors Iain Johnstone, Robert Tibshirani and especially Bradley Efron for many helpful suggestions and comments. He also thanks an anonymous referee for a thorough review that greatly improved the paper. This research was supported by an NSF VIGRE fellowship.

Appendix: Proofs

A.1. Proof of Lemma 1.

Let $\|g\|_h = \left(\int g^2 h dz\right)^{1/2}$. We have

$$\begin{aligned}\Delta(\hat{t}_\rho, G)^{1/2} &= \left\| \frac{\hat{f}'}{\hat{f} \vee \rho} - \frac{f_G'}{f_G} \right\|_{f_G} \\ &\leq \left\| \frac{\hat{f}'}{\hat{f} \vee \rho} - \frac{f_G'}{f_G \vee \rho} \right\|_{f_G} + \left\| \frac{f_G'}{f_G} - \frac{f_G'}{f_G \vee \rho} \right\|_{f_G}.\end{aligned}$$

The second term is the tempering term: $\|f_G'/f_G - f_G'/(f_G \vee \rho)\|_f^2 = \int (1 - f_G/\rho)_+^2 \left(f_G'/f_G\right)^2 f_G dz$. The first term is

$$\begin{aligned}\| \frac{\hat{f}'}{\hat{f} \vee \rho} - \frac{f_G'}{f_G \vee \rho} \|_{f_G} &= \left\| \frac{\left(\hat{f}' - f_G'\right)}{\hat{f} \vee \rho} - \frac{f_G' \left(\hat{f} \vee \rho - f_G \vee \rho\right)}{(f_G \vee \rho)\left(\hat{f} \vee \rho\right)} \right\|_{f_G} \\ &\leq \left\| \frac{\left(\hat{f}' - f_G'\right)}{\hat{f} \vee \rho} \right\|_{f_G} + \left\| \frac{f_G' \left(\hat{f} \vee \rho - f_G \vee \rho\right)}{(f_G \vee \rho)\left(\hat{f} \vee \rho\right)} \right\|_{f_G} \\ &\leq \frac{1}{\rho} \left\| \hat{f}' - f_G' \right\|_{f_G} + \frac{1}{\rho} A(\rho) \left\| \hat{f} \vee \rho - f_G \vee \rho \right\|_{f_G}.\end{aligned}$$

Using $\left|\hat{f} \vee \rho - f_G \vee \rho\right| \leq \left|\hat{f} - f_G\right|$ completes the proof.

A.2. Proof of Theorem 1

Proof. We first bound $\int (\hat{f}' - f_G')^2 dz = \|\hat{f}' - f_G'\|_1^2$ (this is the L^2 norm with

weight 1, not the L^1 norm).

$$\begin{aligned}\left\|\hat{f}' - f_G'\right\|_1^2 &= \frac{1}{2\pi}\int u^2\left(\hat{f}^* - f_G^*\right)^2 du \\ &\leq \frac{1}{2\pi}\left(\int a^2\left(\hat{f}^* - f_G^*\right)^2 du + \int_{|u|\geq a} u^2\left(\hat{f}^* - f_G^*\right)^2 du\right) \\ &= \frac{a^2}{2\pi}\left(\left\|\hat{f} - f_G\right\|_1^2 + \frac{1}{a^2}\int_{|u|\geq a} u^2\left(\hat{f}^* - f_G^*\right)^2 du\right) \\ &\leq \frac{a^2}{2\pi}\left(\left\|\hat{f} - f_G\right\|_1^2 + 4\frac{1}{a^2}\int_{|u|\geq a} u^2 H(u)^2 du\right) \\ &= \frac{a^2}{2\pi}\left(\left\|\hat{f} - f_G\right\|_1^2 + 4L(a)\right)\end{aligned}$$

for all $a \geq C$. We know $L(a) \to 0$ as $a \to \infty$ and L is monotone. Let $a = L^{-1}(\|\hat{f} - f_G\|_1^2)$, or if $L < \|\hat{f} - f_G\|_1^2$, take $a = C$. Then

$$\left\|\hat{f}' - f_G'\right\|_1^2 \leq \frac{5}{2\pi}L^{-1}\left(\left\|\hat{f} - f_G\right\|_1^2\right)^2\left\|\hat{f} - f_G\right\|_1^2.$$

Now plug this bound into Lemma 1. For any function g, we have $\|g\|_{f_G} \leq \|f_G\|_\infty^{1/2}\|g\|_1$. Thus

$$\begin{aligned}\Delta\left(\hat{t}_\rho, G\right)^{1/2} &\leq \frac{1}{\rho}\left\|\hat{f}' - f_G'\right\|_{f_G} + \frac{1}{\rho}A(\rho)\left\|\hat{f} - f_G\right\|_{f_G} + T(\rho, f_G) \\ &\leq \frac{1}{\rho}\|f_G\|_\infty^{1/2}\left(\sqrt{\frac{5}{2\pi}}L^{-1}\left(\left\|\hat{f} - f_G\right\|_1^2\right) + A(\rho)\right)\left\|\hat{f} - f_G\right\|_1 + T(\rho, f_G).\end{aligned}$$

A.3. Proof of Theorem 2

Note that $(\log f_t)' = t + f_0'/f_0$, so $(\log f_G - \log f_t)' = t_G - t$, and $\log f_G/f_t = \int_{z_0}^z (t_G - t)(s)ds$.

The rest of the proof is a simple application of Fubini's theorem. We have

$$\begin{aligned}\int\left|\log\frac{f_G}{f_t}\right| f_G dz &\leq \int\left|\log\frac{f_G}{f_t}\right| f_G dz \\ &\leq \int_{-\infty}^{\infty}\int_{z_0}^{z}|t_G - t|(s) f_G(z) ds dz.\end{aligned}$$

The integrand is positive, so:

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{z_0}^{z} |t_G(s)-t(s)|\, f_G(z)dsdz \\
&\quad = \int_{z_0}^{\infty} |t_G(s)-t(s)|\,(1-F_G(s))\,ds + \int_{-\infty}^{z_0} |t_G(s)-t(s)|\, F_G(s)ds \\
&\quad = \int_{-\infty}^{\infty} |t_G(s)-t(s)|\, P(s) f_G(s)ds.
\end{aligned}$$

Applying Cauchy-Schwartz finishes the proof.

A.4. Proof of Corollary 1

We have

$$\begin{aligned}
|f_G^*(u)| &\leq \int |f_\theta^*(u)|\, dG \\
&= \int \left(1+\frac{4u^2}{\theta^2}\right)^{-k/4} dG \\
&\leq \left(1+\frac{4u^2}{m^2}\right)^{-k/4} + P(\theta \geq m).
\end{aligned}$$

Take $m = u^\alpha$. Then $P(\theta \geq m) = \mathcal{O}\left(u^{-(1-\alpha)k/2}\right)$, and

$$\begin{aligned}
|f_G^*(u)| &\leq \left(1+4u^{2-2\alpha}\right)^{-k/4} + P(\theta \geq m) \\
&= \mathcal{O}\left(u^{(1-\alpha)k/2}\right).
\end{aligned}$$

So for some constants E, M, $|f_G^*(u)|, |\hat{f}^*(u)| \leq Mu^{-(1-\alpha)k/2}$ for all $|u| \geq E$. Thus we can take $H(u) = Mu^{-(1-\alpha)k/2}$. Also, note that $\|f_G\|_\infty \leq \int \|f_\theta\|_\infty dG = E_G(\theta)\|f_1\|_\infty$. Plugging into Theorem 1 completes the proof.

A.5. Proof of Corollary 2

In this proof, C denotes a generic constant, not necessarily the same from line to line. If θ is bounded, we can take $A(\rho) = (1/\rho)\sup_\theta \|f_\theta'\|_\infty \leq C/\rho$. Since $P(\theta \geq m) = 0$ for m large, we can take any $\alpha < 1 - 4/k$. Then by Corollary 1,

$$\Delta\left(\hat{t}_\rho, G\right)^{1/2} \leq C\left[\frac{1}{\rho} d\left(\hat{f}, f_G\right)^{1-2/(1+(1-\alpha)k)} + \frac{1}{\rho^2} d\left(\hat{f}, f_G\right) + T(\rho, f_G)\right].$$

Now we find the order of the tempering term. We have

$$T(\rho, f_G)^2 = \int \left(1 - \frac{f_G}{\rho}\right)_+^2 \left(\frac{f_G'}{f_G}\right)^2 f_G dz$$
$$\leq P_G(f_G(z) < \rho)^{1/2} E_G\left(\left(\frac{f_G'}{f_G}\right)^4\right)^{1/2}$$

so we need to bound $P_G(f_G(z) < \rho)$. If ρ is small, $f_G(z) < \rho$ if x is large or small. Consider the case of large z. For z sufficiently large, $f_G(z) \geq f_{M_1}(z) = C_k z^{k/2-1} M_1^{k/2} \exp(-M_1 z/2)$. Let z_0 be the largest z such that $\rho = f_{M_1}(z_0)$. Then

$$\begin{aligned} P_G(f_G(z) < \rho, z \text{ large}) &\leq P_G(f_{M_1}(z) \leq \rho, z \text{ large}) \\ &= P_{M_2}(f_{M_1}(z) \leq \rho, z \text{ large}) \\ &= P_{M_2}(z \geq z_0) \\ &\approx C z_0^{k/2-1} \exp\left(-M_2 \frac{z_0}{2}\right) \\ &= C\rho \exp((M_1 - M_2) z_0) \\ &\leq C\rho \end{aligned}$$

using the asymptotic expasion $\lim_{x\to\infty} \int_x^\infty t^{s-1} \exp(-t) dt / (x^{s-1} \exp(-x)) = 1$. Similarly $P_G(f_G(z) < \rho, z \text{ small}) \leq C\rho$. So $T(\rho, f_G) = \mathcal{O}(\rho^{1/2})$. Now choose $\rho = \mathcal{O}(\|\hat{f} - f_G\|_1^{2/5})$. Then $\Delta(\hat{t}_\rho, G) = \mathcal{O}(\|\hat{f} - f_G\|_1^{2/5})$.

A.6. Gamma mixture model details

We choose a, b as follows. We first specify a number of groups ℓ. Next, we fit a Gamma prior $\tilde{G}$ by method of moments as for the parametric EB method, and find $\mu = E_{\tilde{G}}(\log\theta)$ and $\sigma = Var_{\tilde{G}}(\log\theta)$. We then take a sequence of means from $\mu - 3\sigma$ to $\mu + 3\sigma$, $\tilde{\mu} = seq(\mu - 3\sigma, \mu + 3\sigma, length = \ell)$ in R notation. Finally, we initialize a, b so each group has approximate log-mean $\tilde{\mu}$ and approximate log-variance $\tilde{\sigma} = (\tilde{\mu}_2 - \tilde{\mu}_1)^2$. To do this, we take $a = 1/\tilde{\sigma}$ and $b = \exp(\psi'(a) - \tilde{\mu})$ where ψ is the digamma function.

For the EM algorithm, we initialize π to be approximately lognormal, $\pi \propto dnorm(\tilde{\mu}, \mu, \sigma)$ in R notation. For the E-step, we estimate

$$g_{ij} = P(z_i \text{ from group } j) = \frac{\pi_j f_j(z_i)}{\sum \pi_j f_j(z_i)},$$

where

$$f_j(x) = \frac{1}{x}\left(\frac{\Gamma(a + k/2)}{\Gamma(k/2)\Gamma(a)}\right)\left(1 - \frac{x/2}{b + x/2}\right)^a \left(\frac{x/2}{b + x/2}\right)^{k/2}$$

is the marginal distribution corresponding to a $Gamma(a, b)$ prior. For the M-step, we estimate $\pi_j \leftarrow (1/N)\sum_i g_{ij}$.

References

Berger, J. (1980). Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters. *Ann. Statist.* **8**, 545-571.

Berger, J. O. and Srinivasan, C. (1978). Generalized Bayes estimators in multivariate problems. *Ann. Statist.* **6**, 783-801.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42**, 855-903.

Brown, L. D. and Greenshtein, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* **37**, 1685-1704.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23**, 1-22.

Efron, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.* **104**, 1015-1028.

Efron, B. and Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators–Part I: The Bayes case. *J. Amer. Statist. Assoc.* **66**, 807-815.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 117-130.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37**, 1647-1684.

Robbins, H. (1954). An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1**, 157-163.

Sharma, D. (1973). Asymptotic equivalence of two estimators for an exponential family. *Ann. Statist.* **1**, 973-960.

Singh, R. S. (1979). Empirical bayes estimation in lebesgue-exponential families with rates near the best possible rate. *Ann. Statist.* **7**, 890-902.

Singh, R. S. and Wei, L. (1992). Empirical Bayes with rates and best rates of convergence in $u(x)C(\theta)\exp(-x/\theta)$-family: Estimation case. *Ann. Inst. Statist. Math.* **44**, 435-449.

Zhang, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statist. Sinica* **7**, 181-193.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods: invited paper. *Ann. Statist.* **31**, 379-390.

Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA.

E-mail: omkar@stanford.edu

(Received January 2010; accepted August 2011)