Poisson Compound and Empirical Bayes Estimation, Revisited[1]

Lawrence Brown
Statistics Department, University of Pennsylvania;
e-mail: lbrown@wharton.upenn.edu

Oberwolfach, March15, 2012
(joint work with E. Greenshtein and Y. Ritov)

We investigate a classical non-parametric Poisson empirical Bayes estimation problem and propose an estimator that performs better than the original proposal of Robbins (1955).

Begin with independent Poisson obervations, $Y_i \sim Po(\lambda_i)$, $i = 1,..,p$, indep. Consider the standard decision theoretic estimation problem. Estimate the vector $\lambda = (\lambda_1,..,\lambda_p)$ by $\delta = \delta(Y)$. Consider the average quadratic risk $R(\delta,\lambda) = E_\lambda\left(\boldsymbol{p}^{-1}\|\delta-\lambda\|^2\right)$. For a prior distribution, *G*, the expected risk is denoted by $R(G,\delta) = E_G\left(R(\Lambda,\delta)\right)$. The Bayes procedure is $\delta_G(y) = E(\Lambda|Y=y)$ (with the conditional expectation taken coordinate-wise). The Bayes risk is $B(G) = R(G,\delta_G) = \min_\delta R(G,\delta)$.

Here is the classical empirical Bayes estimator proposed in Robbins (1955) Here, *G* is unknown. The goal is to find an estimator $\tilde{\delta}$ that approximates $\delta_G$ sufficiently well so that is "*small*" uniformly in $G$ as $p \to \infty$. In this setting, "*small*" can mean *o*(1) or sometimes something even smaller, if possible.

The approach we take is consistent with Robbins original empirical Bayes proposal. Write $\delta_G$ as a functional of the marginal distribution $P_G(y) = \int Po_\lambda(y)G(d\lambda)$. *ie,* $\delta_G = \Delta(P_G)$. Then use the sample $Y = Y_1,..,Y_p$ to estimate $P_G$ by, say, $\tilde{P}$, and $\delta_G$ by $\tilde{\tilde{\delta}} = \Delta(\tilde{P})$. In his paper, Robbins took such an approach. He observed that if *G* is known then the Bayes estimator can be written as

$$\delta_G(y) = \frac{\int \lambda Po(y|\lambda)G(d\lambda)}{\int Po(y|\lambda)G(d\lambda)}$$
$$= \frac{\int (y+1)Po(y+1|\lambda)G(d\lambda)}{\int Po(y|\lambda)G(d\lambda)} = \frac{(y+1)P_G(y+1)}{P_G(y)}.$$

Note that this is a function of the marginal distribution $P_G$. Summarize the observed sample by $\{\mathbb{N}_Y(k)\}$ where $\mathbb{N}_Y(k) = \#\{Y_i : Y_i = k\}$. Then a natural estimator of $P_G$ is $\hat{P}(y) = Z_y/p$. This suggests the following empirical Bayes estimator, which is known as Robbins' estimator for this problem --

$$\hat{\tilde{\delta}}(k) = \frac{(k+1)\hat{P}(k+1)}{\hat{P}(k)} = \frac{(k+1)\mathbb{N}_{Y+1}(k+1)}{\mathbb{N}_Y(k)}.$$

It's clear that for any fixed *G* and each *y*, $\hat{P}(y) \to P_G(y)$ as $p \to \infty$ and $\hat{\tilde{\delta}}(y) \to y$. However, there are some serious problems with $\hat{\tilde{\delta}}$:

1. Problem 1: If $\mathbb{N}(k+1) > 0$ but $\mathbb{N}(k) = 0$ (or is small) then $\hat{\tilde{\delta}}(Y_i = k) = \infty$ (or is probably not desirably accurate).
2. Problem 2: Any Bayes estimator is monotone in *y*; but $\hat{\tilde{\delta}}$ is not.
3. Problem 3 (a subcase of P2). At $y_{(p)} = \max\{y_i\}$ we have $\hat{\tilde{\delta}}(y_{(p)}) = 0$.

The remainder of the construction is devoted to modifying the estimator so as to remedy these problems.

To address Problem 1, pick a small $h > 0$ (called the "corruption" parameter). Let $Q \sim Po(h)$. Choices for *h* in the range $0.5 \le h \le 3$ seem to work well. We'll later propose a cross-validation step to choose *h*. Let $Z = Y + Q$. Use $\mathbb{N}_Y(k)$ as a basis for estimating the marginal distribution of *Z*. The estimate is $\tilde{P}_Z(z) = \sum_{j=0}^{z} \frac{\mathbb{N}(j)}{p} \frac{h^{z-j}e^h}{(z-j)!}$. Now generate $Q_i \sim_{iid} Po(h)$ and build a corrupted sample $\{Z_i\}$ with $Z_i = Y_i + Q_i$. (Each $Z_i \sim_{iid} Po(\lambda_i + \boldsymbol{h})$.) Apply Robbins' method to estimate $\lambda_i$ from this sample via $\tilde{\tilde{\delta}}_{h;1}(z) = \frac{(z+1)\tilde{P}_Z(z+1)}{\tilde{P}_Z(z)} - \boldsymbol{h}$. It is easily checked that $\tilde{\tilde{\delta}}_{h;1}(z) > 0$ for all $z \ge y_{(1)}$. So define $\tilde{\tilde{\delta}}_{h;1}(z) = 0$ for all $z < y_{(1)}$. This guarantees P1 doesn't happen.

However, $\{\tilde{\tilde{\delta}}_{h;1}(z_i)\}$ is a randomized estimator, since $Z_i = Y_i + Q_i$. Such estimators can be improved. To do so, Rao-Blackwellize. Let $\tilde{\tilde{\delta}}_{h;\boldsymbol{2}}(y) = E^Q\left(\hat{\tilde{\delta}}_{h;1}(y+Q)\right) = \sum \frac{h^j e^{-h}}{j!} \tilde{\tilde{\delta}}_{h;1}(y+j)$. The random $Q_i$ have now disappeared. The estimator $\{\tilde{\tilde{\delta}}_{h;\boldsymbol{2}}(y_i)\}$ is a closed-form function of $\{Y_j\}$ through the sufficient statistics $\{\mathbb{N}_Y(k); k = 0,\ldots\}$.

Problem 2 usually persists – $\tilde{\tilde{\delta}}_{h;\boldsymbol{2}}(y)$ need not be monotone in *y*.

S0 we monotone-ize $\tilde{\tilde{\delta}}_{h;\boldsymbol{2}}$. As a convenient, but rather ad-hoc method, we use the Pool-Adjacent-Violators algorithm developed for least-squares isotonic regression. Koenker and Mizra (unpublished) have proposed a more principled and likely better method that appears to still be computationally feasible. It can be verified that so

long as *h* is not too small, this should also fix any remnant of P3. Call the resulting monotone-ized estimator $\Delta_h$.

It remains only to choose the corruption parameter, *h*. One plausible possibility that generally works well on examples is to directly choose a moderate value of *h* – say $1 \le h \le 3$. A more interesting and flexible choice involves what we call "*inbred* cross-validation":

Let $p < 1$ but not too far from 1. Let $B_i \sim_{ind} Bin(Y_i, p)$. Let $U_i = B_i$ and $V_i = Y_i - U_i$. This yields $U_i \sim Po(p\lambda_i)$, $V_i \sim Po((1-p)\lambda_i)$ and $U_i \perp\!\!\!\perp V_i$. Then use $\Delta_h$ on the sample $\{U_i\}$ to estimate $\{p\lambda_i\}$, and use cross-validation on the smaller sample $\{V_i\}$ to choose *h*. The estimates of $\{p\lambda_i\}$ can be adjusted to estimate $\{\lambda_i\}$. It is possible to also use an additional Rao-Blackwell step here to further improve the estimator, but we did not do so for simulations that we have reported.

Asymptotics of Robbins' method are quite appealing. But simulations we have performed show that actual performance in examples can be quite sub-optimal. Here is a slightly informal statement of a theorem we have proved.

If $\{G_k\}$ is a sequence of priors on a bounded set that does not concentrate at a single point then a rate-sharp bound is

$$R(G_k, \hat{\delta}) - B(G_k) = O\left(\frac{(\log p)^2}{\log\log p}\right).$$

This isn't much larger than $(\log p)^2$, and so seems a pretty desirable convergence rate. But behavior in finite (not too large) samples can be much worse than this suggests, as revealed by simulations we have performed for a variety of examples. For *p* = 200 and *G* supported within [0,20], Robbins' estimator can be worse than $\Delta_h$ by 5 – 35% in terms of squared error risk, depending on the form of *G*.

References

Robbins, H. (1955). An Empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. on Prob. Statist.* 157-164.