Bayesian statistics and the borrowing of strength in high-dimensional data analysis

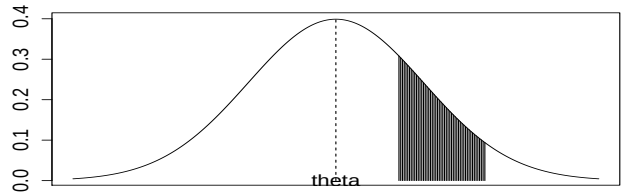
Aad van der Vaart
Mathematical Institute
Leiden University

Royal Netherlands Academy of Sciences, Amsterdam, September 2013

Measurements with errors
(Gauss, 1810, Fisher, 1930, Cramer, Rao, 1950, Lehmann, 1970)

AIM: determine θ

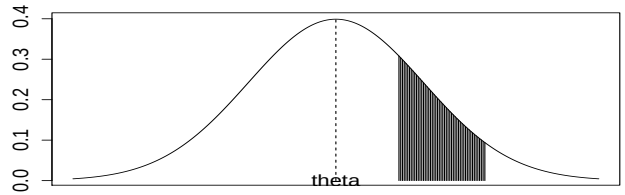
MEASUREMENTS: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$.



Measurements with errors
(Gauss, 1810, Fisher, 1930, Cramer, Rao, 1950, Lehmann, 1970)

AIM: determine θ

MEASUREMENTS: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$.



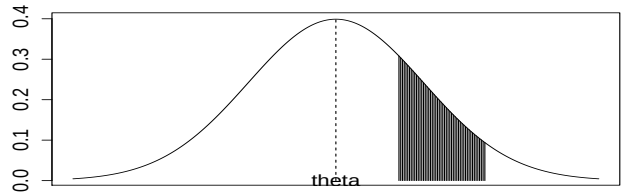
Optimal method to recover θ:

$$\hat{\theta} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Measurements with errors (Gauss, 1810, Fisher, 1930, Cramer, Rao, 1950, Lehmann, 1970)

AIM: determine θ

MEASUREMENTS: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$.



Optimal method to recover θ:

$$\hat{\theta} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Principles:

- Maximum likelihood
- Objective Bayes
- Equivariant

Criteria:

- Minimum variance unbiased
- Admissible for symmetric loss
- Minimal risk equivariant
- Minimax

Multidimensional measurements with errors (Stein, 1956)

AIM: determine $\theta^1, \ldots, \theta^p$

MEASUREMENTS: $X_1^j, \ldots, X_n^j \overset{\text{iid}}{\sim} N(\theta^j, 1)$, for $j = 1, \ldots, p$.

We assume NOT:

- relations between $\theta^1, \ldots, \theta^p$.
- dependence between measurements.

Multidimensional measurements with errors (Stein, 1956)

AIM: determine $\theta^1, \ldots, \theta^p$

MEASUREMENTS: $X_1^j, \ldots, X_n^j \overset{\text{iid}}{\sim} N(\theta^j, 1)$, for $j = 1, \ldots, p$.

We assume NOT:

- relations between $\theta^1, \ldots, \theta^p$.
- dependence between measurements.

THEOREM [Stein, 1956]
If $p \geq 3$, then $(\overline{X_n^1}, \overline{X_n^2}, \ldots, \overline{X_n^p})$ is *inadmissible*:
there exists $(T^1, \ldots, T^p)$ with, for all $\theta^1, \ldots, \theta^p$:

$$\sum_{j=1}^{p} \mathrm{E}(T^j - \theta^j)^2 < \sum_{j=1}^{p} \mathrm{E}(\overline{X_n^j} - \theta^j)^2.$$

Intermezzo: Bayes's rule (Bayes, 1763)

BAYES'S RULE

If a variable θ follows a probabiity distribution Π and given θ a variable X follows a probability density $x \mapsto p(x|\,\theta)$, then θ given X follows the distribution:

$$d\Pi(\theta|\,X) \propto p(X|\,\theta)\, d\Pi(\theta).$$

Intermezzo: Bayes's rule (Bayes, 1763)

BAYES'S RULE

If a variable θ follows a probabiity distribution Π and given θ a variable X follows a probability density $x \mapsto p(x|\,\theta)$, then θ given X follows the distribution:

$$d\Pi(\theta|\,X) \propto p(X|\,\theta)\, d\Pi(\theta).$$

Bayesian statistics:

- $d\Pi(\cdot)$ models the a-priori uncertainty about the parameter θ.
- $d\Pi(\cdot|\,X)$ the a-posteriori uncertainty.

Empirical Bayes method (Robbins, 1960s)

WORKING HYPOTHESIS: $(\theta^1, \theta^2, \ldots, \theta^p) \sim \Pi$.

Empirical Bayes method (Robbins, 1960s)

WORKING HYPOTHESIS: $(\theta^1, \theta^2, \ldots, \theta^p) \sim \Pi$.

Then $(X_i^j, \theta^j : i = 1, \ldots, n, j = 1, \ldots, p)$ follow a joint probability distribution.

Bayes's rule gives a conditional distribution

$$(\theta^j : j = 1, \ldots, p) \,|\, (X_i^j : i = 1, \ldots, n, j = 1, \ldots, p)$$

and hence expected values

$$T^j(\Pi) := \mathrm{E}(\theta^j \,|\, X_i^j : i = 1, \ldots, n, j = 1, \ldots, p).$$

Estimate Π from the data and use $T^j(\hat{\Pi})$ (or use a hyper prior).

Empirical Bayes method: example (James, Stein)

AIM: determine $\theta^1, \ldots, \theta^p$

MEASUREMENTS: $X_1^j, \ldots, X_n^j \overset{\text{iid}}{\sim} N(\theta^j, 1)$, for $j = 1, \ldots, p$.

WORKING HYPOTHESIS: $\theta^1, \ldots, \theta^p \overset{\text{iid}}{\sim} N(0, A)$.

If we knew A, then we might use the *Bayes estimator* $A/(A+1)\overline{X_n^j}$.

Empirical Bayes method: example (James, Stein)

AIM: determine $\theta^1, \ldots, \theta^p$

MEASUREMENTS: $X_1^j, \ldots, X_n^j \overset{\text{iid}}{\sim} N(\theta^j, 1)$, for $j = 1, \ldots, p$.

WORKING HYPOTHESIS: $\theta^1, \ldots, \theta^p \overset{\text{iid}}{\sim} N(0, A)$.

If we knew A, then we might use the *Bayes estimator* $A/(A+1)\overline{X_n^j}$.

Under the working hypothesis $\overline{X_n^j} \overset{\text{iid}}{\sim} N(0, A + 1/n)$.

This suggests the estimate $\hat{A} = \sum_j \overline{X_n^j}^2/(p-2) - 1/n$.

Empirical Bayes method: example (James, Stein)

AIM: determine $\theta^1, \ldots, \theta^p$

MEASUREMENTS: $X_1^j, \ldots, X_n^j \overset{\text{iid}}{\sim} N(\theta^j, 1)$, for $j = 1, \ldots, p$.

WORKING HYPOTHESIS: $\theta^1, \ldots, \theta^p \overset{\text{iid}}{\sim} N(0, A)$.

If we knew A, then we might use the *Bayes estimator* $A/(A+1)\overline{X_n^j}$.

Under the working hypothesis $\overline{X_n^j} \overset{\text{iid}}{\sim} N(0, A + 1/n)$.

This suggests the estimate $\hat{A} = \sum_j \overline{X_n^j}^2/(p-2) - 1/n$.

$\hat{A}/(\hat{A}+1)\overline{X_n^j}$ beats the MLE.

Not only under the working hypothesis, but for *any* $\theta^1, \ldots, \theta^p$.

Nonparametric empirical Bayes method (Zhang, 2009)

AIM: determine $\theta^1, \ldots, \theta^p$

MEASUREMENTS: $X_1^j, \ldots, X_n^j \overset{\text{iid}}{\sim} N(\theta^j, 1)$, for $j = 1, \ldots, p$.

WORKING HYPOTHESIS: $\theta^1, \ldots, \theta^p \overset{\text{iid}}{\sim} G$.

Then $\overline{X_n^j}$ has marginal density $x \mapsto \int \sqrt{n}\phi\big((x-s)\sqrt{n}\big)\, dG(s)$, and *nonparametric maximum likelihood estimator* for G is:

$$\hat{G} = \operatorname*{argmax}_G \prod_{j=1}^{p} \int \sqrt{n}\phi\big((\overline{X_n^j} - s)\sqrt{n}\big)\, dG(s).$$

Nonparametric empirical Bayes method (Zhang, 2009)

AIM: determine $\theta^1, \ldots, \theta^p$

MEASUREMENTS: $X_1^j, \ldots, X_n^j \overset{\text{iid}}{\sim} N(\theta^j, 1)$, for $j = 1, \ldots, p$.

WORKING HYPOTHESIS: $\theta^1, \ldots, \theta^p \overset{\text{iid}}{\sim} G$.

Then $\overline{X_n^j}$ has marginal density $x \mapsto \int \sqrt{n}\phi\big((x-s)\sqrt{n}\big)\, dG(s)$, and *nonparametric maximum likelihood estimator* for G is:

$$\hat{G} = \operatorname*{argmax}_G \prod_{j=1}^{p} \int \sqrt{n}\phi\big((\overline{X_n^j} - s)\sqrt{n}\big)\, dG(s).$$

(An analogous *full Bayes analysis* would put a Dirichlet prior on G.)

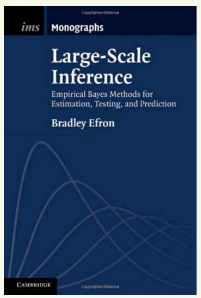
Borrowing strength

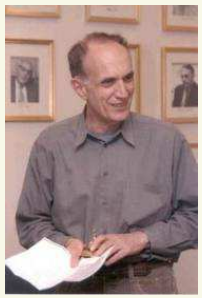
Borrowing strength: by connecting (even seemingly unrelated) samples and parameters we gain overall.

Borrowing strength

Borrowing strength: by connecting (even seemingly unrelated) samples and parameters we gain overall.

This is particularly important in *large scale inference*, when the data is massive, as in genomics, systems biology, image analysis, ...





Bradley Efron

Does it work?

Bayesian assumptions on parameters are often *working hypotheses*, not based on scientific theory.

Frequentist Bayesian theory tries to validate (or not) the resulting procedures in a general, non-Bayesian framework, taking account that priors can be partly *misspecified*.

> Bayesian methods are promising for high-dimensional data, but their performance is poorly understood at the present time.

History

Stein's 1956 *inadmissibility* seemed a peculiarity.

History

Stein's 1956 *inadmissibility* seemed a peculiarity.

Robbins' 1960s *empirical Bayes method* was not followed up much either.

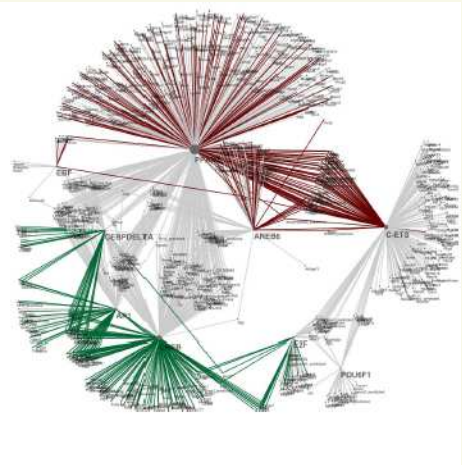
History

Stein's 1956 *inadmissibility* seemed a peculiarity.

Robbins' 1960s *empirical Bayes method* was not followed up much either.

OUTLOOK:

- In high dimensions the potential gain is large.
- A-priori knowledge should make this gain even bigger.



From PhD thesis Geert Geeven, 2010.

Sparsity

Consider many parameters $\theta^1, \theta^2, \ldots, \theta^p$, but suppose most are actually (near) zero.

Sparsity prior:

- Choose s from prior π_n on $\{0, 1, 2, \ldots, p\}$ with exponential decrease.
- Choose $S \subset \{0, 1, \ldots, p\}$ of size $|S| = s$ at random.
- Choose $(\theta_i : i \in S)$ from density g_S on $\mathbb{R}^S$ (and set other θ_i zero).

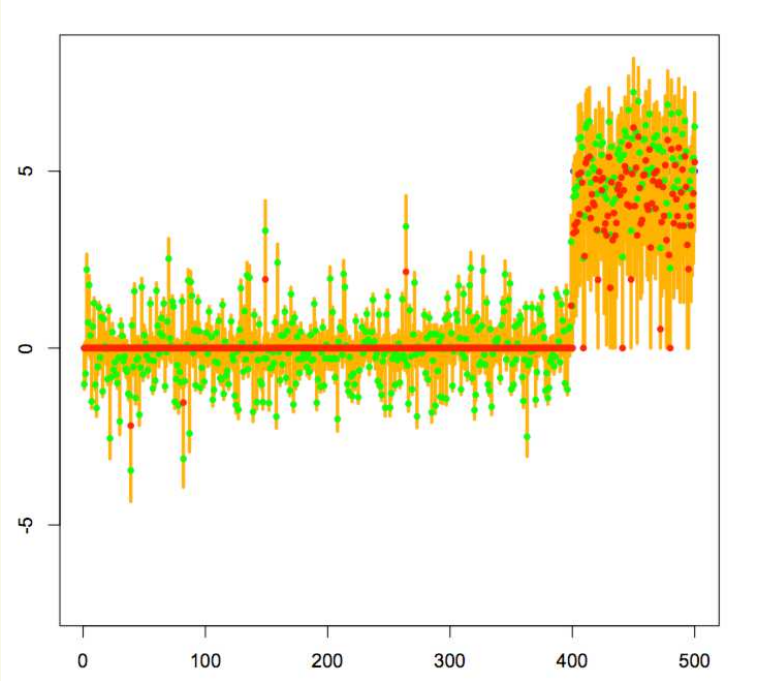
THEOREM [Castillo, vdV, 2013]
This achieves the *minimax benchmark*: for $s = \#(j : \theta^j \neq 0)$,

$$\mathrm{E} \sum_{j=1}^{p} (T^j - \theta^j)^2 \asymp \frac{s}{n} \sqrt{\log \frac{s}{n}}.$$

Compare to: $\mathrm{E} \sum_{j=1}^{p} (\overline{X_n^j} - \theta^j)^2 = \frac{p}{n}$.

Uncertainty quantification and multiplicity correction

The Bayesian analysis results in a posterior distribution on $(\theta^1, \ldots, \theta^p)$, and hence in marginal posterior distributions of every θ^j.



Marginal credible intervals (orange)

Credible intervals can be used for overall uncertainty quantification (?).

Genomics: RNA sequencing (vd Wiel, Leday, Rue, v Wieringen, vdV, *Biostatistics, 2013)*

$Y_{i,j}$: RNA expression count of tag $i = 1, \ldots, p$ in tissue $j = 1, \ldots, n$.

x_j: covariates of tissue j.

$Y_{i,j} \sim$ (zero-inflated) *negative binomial*, with

$$\mathrm{E}Y_{i,j} = e^{\alpha_i + \beta_i x_j}, \qquad \mathrm{var}\, Y_{i,j} = \mathrm{E}Y_{i,j}\left(1 + \mathrm{E}Y_{i,j} e^{-\phi_i}\right).$$

Simple Bayesian model: $\alpha_i \perp \beta_i \perp \phi_i$ with

$$\alpha_i \sim F, \qquad \beta_i \sim G_\tau, \qquad \phi_i \sim H_\tau.$$

Genomics: RNA sequencing (vd Wiel, Leday, Rue, v Wieringen, vdV, *Biostatistics, 2013)*

$Y_{i,j}$: RNA expression count of tag $i = 1, \ldots, p$ in tissue $j = 1, \ldots, n$.

x_j: covariates of tissue j.

$$Y_{i,j} \sim \text{(zero-inflated) } \textit{negative binomial}\text{, with}$$

$$\mathrm{E}Y_{i,j} = e^{\alpha_i + \beta_i x_j}, \qquad \operatorname{var} Y_{i,j} = \mathrm{E}Y_{i,j}\left(1 + \mathrm{E}Y_{i,j} e^{-\phi_i}\right).$$

Simple Bayesian model: $\alpha_i \perp \beta_i \perp \phi_i$ with

$$\alpha_i \sim F, \qquad \beta_i \sim G_\tau, \qquad \phi_i \sim H_\tau.$$

Efficient *empirical Bayes approach* to estimate τ: calculate *marginal* posteriors Π^{β_i} of $\beta_1, \ldots, \beta_p$ and Π^{ϕ_i} of $\phi_1, \ldots, \phi_p$ *given* τ and equate

$$\frac{1}{p}\sum_{i=1}^{p} \Pi^{\beta_i}(\cdot \,|\, Y_{i1}, \ldots, Y_{i,n}, \tau) = G_\tau(\cdot), \quad \frac{1}{p}\sum_{i=1}^{p} \Pi^{\phi_i}(\cdot \,|\, Y_{i1}, \ldots, Y_{i,n}, \tau) = H_\tau(\cdot).$$

Summary

Even seemingly unrelated analyses can *borrow strength* from each other.

Gains can be particularly big if the data is big.

Bayesian thinking provides methods that can achieve this.

These methods may be computationally challenging.

There is much to be learned about the validity of these methods, in particular their uncertainty quantification.

Large scale testing by empirical Bayes

For every out of 30 000 genes test statistically whether its expression differs in cancer and normal tissues.

Large scale testing by empirical Bayes

For every out of 30 000 genes test statistically whether its expression differs in cancer and normal tissues.

Are the 30 000 tests connected?

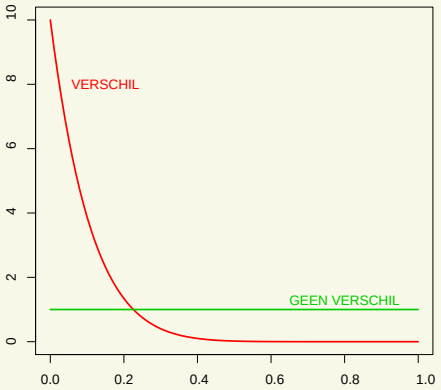
Large scale testing by empirical Bayes

For every out of 30 000 genes test statistically whether its expression differs in cancer and normal tissues.

Are the 30 000 tests connected?

Assume:

- A gene is expressed with probability π.
- The p-value of the gene's test is random if not expressed; otherwise from some f.



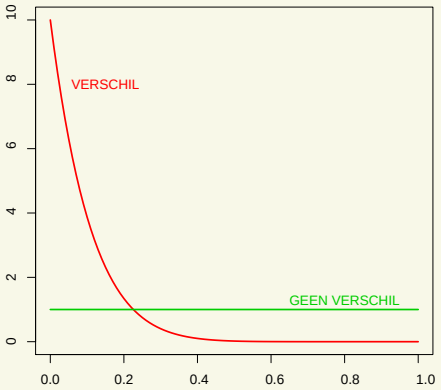
Large scale testing by empirical Bayes

For every out of 30 000 genes test statistically whether its expression differs in cancer and normal tissues.

Are the 30 000 tests connected?

Assume:

- A gene is expressed with probability π.
- The p-value of the gene's test is random if not expressed; otherwise from some f.



Now estimate π and f from the data and compute

$$\mathrm{P}\big(\text{gene is expressed}\big|\,\text{ALL DATA}\big).$$

Bernstein-von Mises theorem

THEOREM

Given 'flat priors' on the $s_n \ll p$ nonzero coefficients,

$$\mathrm{E}_{\theta_0} \left\| \Pi_n(\cdot \,|\, Y^n) - \sum_S \hat{w}_S N(\hat{\theta}_{(S)}, \Gamma_S^{-1}) \otimes \delta_{S^c} \right\| \to 0,$$

for $\hat{\theta}_{(S)}$ the LS estimator for model S, Γ_S^{-1} its covariance, and

$$\hat{w}_S \propto \frac{\pi_p(s)}{\binom{p}{s}} \left(\tfrac{\lambda\sqrt{2\pi}}{2}\right)^s |\Gamma_S|^{-1/2} e^{\frac{1}{2}\|X_S \widehat{\theta}_{(S)}\|_2^2} 1_{|S| \le 4s_n, \|\theta_{0,S^c}\|_1 \lesssim s_n \sqrt{\log p}/\|X\|}.$$

COROLLARY

Given consistent model selection, mixture can be replaced by $N(\hat{\theta}_{(S_{\theta_0})}, \Gamma_{S_{\theta_0}}^{-1})$.