Perspectives in Statistical Modeling and Inference

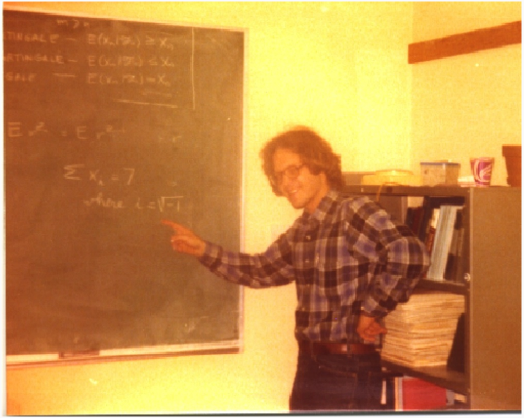
A Workshop in Honor of Ed George's 70th Birthday

Data-dependent priors

Wharton University of Pennsylvania

Nancy Reid

University of Toronto

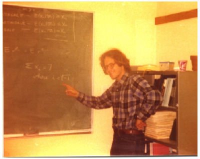


Σ X_i = 7
where i = √-1


ED
GEORGE


MAXIMUM
OCCUPANCY
CAPACITY
(297)

J. Appl. Prob. **24**, 557–573 (1987)
Printed in Israel



SAMPLING RANDOM POLYGONS

EDWARD I. GEORGE,* *University of Chicago*

Abstract

Every realization of a Poisson line process is a set of lines which subdivides the plane into a population of non-overlapping convex polygons. To explore the unknown statistical features of this population, an alternative stochastic construction of random polygons is developed. This construction, which is based on an alternating sequence of random angles and side lengths, provides a fast simulation method for obtaining a random sample from the polygon population. For the isotropic case, this construction is used to obtain a random sample of 2500000 polygons, providing the most precise estimates to date of some of the unknown distributional characteristics.

GEOMETRIC PROBABILITY; POISSON LINE PROCESS

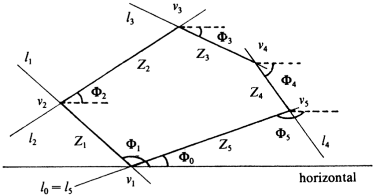


Figure 2.2. The notation for a polygon when $N = 5$

TABLE 1c
Distribution estimates for A

$a =$	0.005	0.010	0.025	0.050	0.100
$P(A \leqq a) =$	0.04536	0.06388	0.09968	0.1392	0.1931
$a =$	0.250	0.500	0.750	1.00	1.50
$P(A \leqq a) =$	0.2924	0.3926	0.4615	0.5140	0.5929
$a =$	2.50	5.00	7.50	10.0	12.5
$P(A \leqq a) =$	0.6944	0.8228	0.8846	0.9201	0.9424
$a =$	15.0	20.0	30.0	50.0	100.0
$P(A \leqq a) =$	0.9574	0.9752	0.9902	0.9978	0.9999

Note: 334 polygons with $A > 10$ were observed.

† The simulation was run on a PDP 10/KI computer using the SAIL programming language. The uniform standard deviates were obtained from the random number generator RAN. Polygons were processed at a rate of 8745 polygons per minute of CPU time.

Ed the frequentist

The Annals of Statistics
1986, Vol. 14, No. 1, 188–205

MINIMAX MULTIPLE SHRINKAGE ESTIMATION

BY EDWARD I. GEORGE

University of Chicago

For the canonical problem of estimating a multivariate normal mean under squared-error-loss, this article addresses the problem of selecting a minimax shrinkage estimator when vague or conflicting prior information suggests that more than one estimator from a broad class might be effective. For this situation a new class of alternative estimators, called multiple shrinkage estimators, is proposed. These estimators use the data to emulate the behavior and risk properties of the most effective estimator under consideration. Unbiased estimates of risk and sufficient conditions for minimaxity are provided. Bayesian motivations link this construction to posterior means of mixture priors. To illustrate the theory, minimax multiple shrinkage Stein estimators are constructed which can adaptively shrink the data towards any number of points or subspaces.

Ed the empirical Bayesian

Biometrika (2000), **87**, 4, *pp.* 731–747

Printed in Great Britain

Calibration and empirical Bayes variable selection

By EDWARD I. GEORGE

Department of Management Science and Information Systems, The University of Texas at Austin, Austin, Texas 78712-1175, U.S.A.

egeorge@mail.utexas.edu

and DEAN P. FOSTER

Department of Statistics, The Wharton School of The University of Pennsylvania, Philadelphia, Pennsylvania 19104-6302, U.S.A.

foster@diskworld.wharton.upenn.edu

Ed the Bayesian

Statistics and Computing (2000) **10**, 17–24

Hierarchical priors for Bayesian CART shrinkage

HUGH CHIPMAN[1], EDWARD I. GEORGE[2] and ROBERT E. McCULLOCH[3]

[1]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1 (hachipman@uwaterloo.ca)*
[2]*Department of MSIS, University of Texas, Austin, TX 78712-1175 (egeorge@mail.utexas.edu)*
[3]*Graduate School of Business, University of Chicago, IL 60637 (Robert.McCulloch@gsbpop.uchicago.edu)*

Submitted March 1998 and accepted March 1999

The Variable Selection Problem

Edward I. GEORGE

The problem of variable selection is one of the most pervasive model selection problems in statistical applications. Often referred to as the problem of subset selection, it arises when one wants to model the relationship between a variable of interest and a subset of potential explanatory variables or predictors, but there is uncertainty about which subset to use. This vignette reviews some of the key developments that have led to the wide variety of approaches for this problem.

"freqentist justification is needed for Bayesian procedures"

JASA 2000 vignette series

Ed the (almost) nonparametric Bayesian

Spike-and-Slab Meets LASSO: A Review of the Spike-and-Slab LASSO *

Ray Bai[†], Veronika Ročková[‡], Edward I. George[§]

May 11, 2021

Ed the calibrated Bayesian

We have from (5.2) that the $100(1-\alpha)\%$ asymptotic pointwise confidence intervals for $\beta_j, j = 1, \ldots, p$, are

$$[\widehat{\beta}_{dj} - c(\alpha, n, \widehat{\sigma}^2), \widehat{\beta}_{dj} + c(\alpha, n, \widehat{\sigma}^2)], \tag{5.3}$$

where $c(\alpha, n, \widehat{\sigma}^2) := \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{\sigma}^2(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}^T)_{jj}/n}$ and $\Phi(\cdot)$ denotes the cumulative distribution function of $\mathcal{N}(0,1)$.

Objective priors

Kass Wasserman 96; Consonni et al. 18

- don't influence the posterior very much
- that we can all agree on
- lead to calibrated posterior credible sets
- anything that modifies the likelihood function

- flat, uniform, vague, highly dispersed, ...
- reference, other minimum information versions
- matching priors
- fiducial, generalized fiducial, default

These are all necessarily model-dependent

Some are data-dependent

Model dependence

- parametric model $f(y;\theta)$, $\quad y \in \mathbb{R}^n; \theta \in \mathbb{R}^p$, $\quad p < n$
- Jeffreys' invariant prior

$$\pi_J(\theta) \propto |i(\theta)|^{1/2}, \qquad i(\theta) = \mathrm{E}\{-\partial^2 \ell(\theta;y)/\partial\theta\partial\theta^T\}; \quad \ell(\theta;y) = \log f(y;\theta)$$

- invariant to reparametrization
- if $p = 1$ $\pi_J(\theta)$ is a matching prior, and a reference prior, and ...
- a Jeffreys'-like prior for $p > 1$ is

$$\pi(\theta) \propto g(\lambda) i_{\psi\psi}(\theta)^{1/2}, \qquad \theta = (\psi, \lambda), \qquad i(\theta) \text{ partitioned }, \qquad \psi \perp \lambda$$

- objective priors need to be targetted on the function of interest

Approximate matching priors

- matching priors ensure calibration of confidence bounds
- posterior credible bound

$$\mathrm{pr}\{\theta \le \theta^{1-\alpha}(y) \mid y\} = 1 - \alpha$$

$$\mathrm{pr}\{\theta^{1-\alpha}(Y) \ge \theta \mid \theta\} = 1 - \alpha + O(n^{-1})$$

- when $p = 1$ matching to $O(n^{-1})$ achieved by Jeffreys' prior

$$\pi_J(\theta) \propto i(\theta)^{1/2}$$

- matching to $O(n^{-3/2})$ only if

$$\frac{d}{d\theta}\left[\mathrm{E}\{\ell'(\theta)^3\} \Big/ i^{3/2}(\theta)\right] = 0$$

model criterion

Data-dependent priors

- Example: transformed regression

$$y^{\lambda} = X\beta + \sigma\epsilon, \qquad \pi(\lambda) = \frac{\pi_0(\lambda)}{\{J(\lambda; y)\}^{(n-p)/n}} \qquad J(y;\lambda) = \prod_{i=1}^{n} \left| \frac{dy_i^{\lambda}}{dy_i} \right|$$

Box & Cox 1964

- Example: mixture model

$$y_i \sim \sum_{j=1}^{k} p_j \phi\{(y_j - \mu_j)/\sigma_j\}, \quad i = 1, \ldots, n; \qquad \pi_n(\theta) = \pi(\theta)\left\{1 - \frac{L_n(\theta)}{\Delta_n(\theta)}\right\}$$

"the only priors that produce intervals with second-order correct coverage are data-dependent"

Wasserman 2000

… data-dependent priors: BFF

- anything that modifies the likelihood function

- Example: the fiducial density (as defined by Fisher) is of the form

$$df = -\frac{\partial}{\partial\theta}F(y;\theta)d\theta = -\frac{\partial}{\partial\theta}F(y;\theta)\frac{f(y;\theta)}{f(y;\theta)} = \underbrace{L(\theta;y)}_{\text{likelihood}}\underbrace{\left|\frac{dy}{d\theta}\right|}_{\text{“prior”}}$$

y fixed at observed value; total derivative for fixed quantile of y

- Example: generalized fiducial density

$$r(\theta;y) \propto \underbrace{f(y;\theta)}_{\text{Likelihood}}\underbrace{J(\theta;y)}_{\text{“prior”}} \qquad J(\theta;y) = D\left\{\frac{d}{d\theta}G(u;\theta)\big|_{u=G^{-1}(y;\theta)}\right\}$$

Hannig 2009ff

- distribution for θ: posterior, confidence, fiducial all of the same form

$$\pi(\theta \mid y) = \frac{f(y;\theta)\pi(\theta)}{m(y)}$$

- version 1: use special model properties to estimate $m(y)$ or its moments
- e.g. Robbins $\mathrm{E}(\theta \mid y) = (y+1)\frac{m(y+1)}{m(y)}$; $\hat{m}(\cdot)$ via multinomial sometimes density estimate
- version 2: $\pi(\theta \mid \alpha)$ hyperparameter $\longrightarrow m(y;\alpha)$, estimate α by maximum likelihood

marginal ML

- e.g. species

$$E(t) = S\int e^{-\theta}(1-e^{-\theta t})\pi(\theta)d\theta \longrightarrow \widehat{E}(t) = S\int e^{-\theta}(1-e^{-\theta t})\widehat{\pi}(\theta)d\theta$$

$\pi(\theta \mid \hat{\alpha}) = \hat{\pi}(\theta)$

- e.g. multiple shrinkage $\delta_* = \sum \rho_k(y)m_k(y) = E_{\pi_*}(\theta \mid y)$, $\rho_k(y) = \mathrm{pr}(\pi_k \mid y)$ George 1986

- "there are good reasons one might want an estimator of $g(\theta)$, involving questions that can't be answered directly in terms of the marginal density" $g = \pi$
- "Taking $\hat{g}$ literally allows for Bayes estimates, e.g. $\widehat{\mathrm{pr}}(\theta \geq 1 \mid z = 3)$"

- can we "take $\hat{g}(\theta)$ literally"? Efron has a particular construction of $\hat{g}$

- the examples above have $\theta_i \sim \pi(\theta \mid \alpha) \longrightarrow f(y_i \mid \theta_i), \quad i = 1, \ldots, n$ $\theta_i \in \mathbb{R}$
- difficult to see if marginalization paradoxes might arise

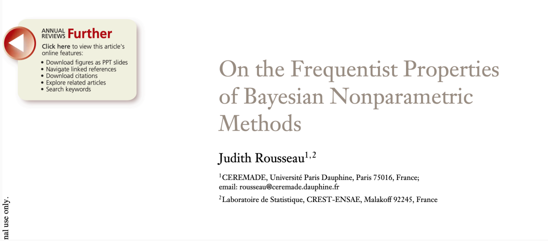
- a "good" prior for θ has unsatisfactory performance for a specific parameter of interest $\psi(\theta)$" Consonni et al.

… empirical Bayes

- how to assess empirical Bayes posteriors such as Efron's g-estimate?
- "Bayes and empirical Bayes: do they merge?" Petrone, Rousseau, Scricciolo (2014)
- $\int \pi(\theta \mid \alpha, y)\pi(\alpha)d\alpha \longrightarrow \pi(\theta \mid \hat{\alpha}_n, y)$ α hyperparameter
- shows that the two methods will agree ("merge") in the limit if EB is consistent

  and much more

- empirical Bayes methods in George & Foster (2000), Cui & George (2008) asymptotic discrepancy in Scott & Berger (2010)



ANNUAL REVIEWS Further

Click here to view this article's online features:

- Download figures as PPT slides
- Navigate linked references
- Download citations
- Explore related articles
- Search keywords

On the Frequentist Properties of Bayesian Nonparametric Methods

Judith Rousseau[1,2]

[1]CEREMADE, Université Paris Dauphine, Paris 75016, France; email: rousseau@ceremade.dauphine.fr

[2]Laboratoire de Statistique, CREST-ENSAE, Malakoff 92245, France

nal use only.

Spike and slab

Spike-and-Slab Meets LASSO: A Review of the Spike-and-Slab LASSO *

Ray Bai[†], Veronika Ročková[‡], Edward I. George[§]

May 11, 2021

- high-dimensional regression $y = X\beta + \epsilon$
-

$$\pi(\beta_j \mid \gamma) = \prod_{j=1}^{p} \{(1-\gamma_j)\psi(\beta_j \mid \lambda_0) + \gamma_j \psi(\beta_j \mid \lambda_1)\}$$

$$\pi(\gamma \mid \alpha) = \prod_{j=1}^{p} \{\alpha^{\gamma_j}(1-\alpha)^{1-\gamma_j}\}$$

$$\alpha \sim Beta(a, b)$$

- adaptive: large amount of shrinkage if $|\beta_j|$ is small or a very small amount of shrinkage if $|\beta_j|$ is large
- borrows strength: marginal prior for β is not a product
- posterior mode can be de-biased

$\hat{\Theta}$ estimated

$$\hat{\beta}_d = \hat{\beta} + \hat{\Theta} X^T (y - X\hat{\beta})/n$$

-

$$\sqrt{n}(\hat{\beta}_d - \beta) \dot{\sim} N(0, \sigma^2 \hat{\Theta}\hat{\Sigma}\hat{\Theta})$$

- leading to confidence intervals for components of β

We have from (5.2) that the $100(1-\alpha)\%$ asymptotic pointwise confidence intervals for $\beta_j, j = 1, \ldots, p$, are

$$[\widehat{\beta}_{dj} - c(\alpha, n, \widehat{\sigma}^2), \widehat{\beta}_{dj} + c(\alpha, n, \widehat{\sigma}^2)], \tag{5.3}$$

where $c(\alpha, n, \widehat{\sigma}^2) := \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{\sigma}^2(\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Theta}}^T)_{jj}/n}$ and $\Phi(\cdot)$ denotes the cumulative distribution function of $\mathcal{N}(0,1)$.

A somewhat random walk

- Donnet et al. (2018) Posterior concentration rates for empirical Bayes procedures
  nonparametric data-dependent priors
- Rousseau & Szabo (2020) Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors
- Martin & Walker (2019) Data-driven priors and their posterior concentration rates
- Zhang & Gao (2020) Convergence rates of empirical Bayes posterior distributions: a variational perspective
- Klebanov et al. (2020) Objective priors in the empirical Bayes framework scalar parameter

A work in progress

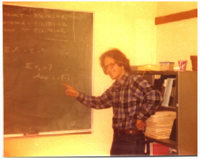
- data-dependent priors are necessary in non-parametric problems
- data-dependent priors are necessary to obtain strong matching in parametric problems
- many interesting parametric models have either $p = p_n$ or $p > n$
- prototype the sequence model $Y_i = \theta_i + \epsilon_i, \quad \epsilon_i \sim N(0, 1/n)$ — or 1, or σ^2/n
- lots of difficult analysis, but what about interpretation?

Model selection and inference

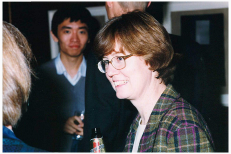
- inference about models is quite difficult
- has much in common with nonparametric methods
- inference after model selection is also very difficult
- Battey & Cox (2017, 2018, 2019) consider finding sets of models that are equally useful

  using ideas from incomplete block designs
- Battey & R (2021) consider inference for individual components β_j without correction for selection

can be used to narrow down selected sets from Battey/Cox strategy



THANK YOU ED



References i

Bai, R., Rockova, V. and George, E.I. (2021). Spike-and-slab meets Lasso: a review of the spike-and-slab Lasso.
`https://arxiv.org/abs/2010.06451`

Box, G.E.P. and Cox, D.R. (1964). Analysis of transformations. *J. R. Statist. Soc.* B **26**, 211–252.

Chipman, H., George, E.I. and McCulloch, R.E. (2000). Hierarchical priors fo Bayesian CART shrinkage. *Statis. Comput.* **10**, 17–24.

Consonni, G., Fouskakis, D., Liseo, B. and Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis* **13**, 627–679.

Cui, W. and George, E.I. (2008). Empirical Bayes vs. fully Bayes variable selection. *J. Statist. Plann. Infer.* **4**, 888–900.

Donnet, S., Rivoirard, V., Rousseau, J. and Scricciolo,C. (2018). Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures. *Bernoulli* **24**, 231–256.

Efron, B. (2021). Empirical Bayes: Concepts and Methods. for *BFF Handbook*.

Efron, B. (2019). Bayes, oracle Bayes and empirical Bayes. *Statist. Sci.* **34**, 177–301.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference.*(Ch. 7). Cambridge University Press.

George, E.I. (1987). Sampling random polygons. *J. Appl. Prob* **24**, 557–573.

George, E.I. (1986). Minimax multiple shrinkage estimation. *Ann. Statist.* **14**, 188–205.

George, E.I. and Foster, D.P. (2000)/ Calibration and empirical Bayes selection. *Biometrika* **87**, 731–747.

References iii

George, E.I. (2000). The variable selection problem. *J. Am. Statist. Assoc.* **95**, 1304–1308.

Hannig, J. (2009). On generalized fiducial inference. *Statist. Sinica* **19**, 491–544.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Am. Statist. Assoc.* **91**, 1343–1370.

Klebanov, I., Sikorski, A., Schütte, C. and Röblitz, S. (2020). Objective priors in the empirical Bayes framework. *Scand. J. Statist.* **48**, 1212–1233. `DOI:10.1111/sjos.12485`.

Martin, R. and Walker, S.G. (2019). Data-driven priors and their posterior concentration rates. *Electron. J. Statist.* **13**, 3049–3081.

Petrone, S., Rousseau, J. and Scricciolo, C. (2014). Bayes and empirical Bayes: do they merge?. *Biometrika* **101**, 285–302.

References iv

Rousseau, J. (2016). On the frequentist properties of Bayesian nonparametric methods. *Ann. Rev. Statist. Appl.* **3**, 211–231.

Rousseau, J. and Szabo, B. (2020). Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. *Ann. Statist.* **48**, 2155–2179.

Scott, J.G. and Berger, J.O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38**, 2587–2619.

Wasserman, L. (2000). Asymptotic inference for mixutre models using data-dependent priors. *J. R. Statist. Soc.* B **62**, 159–180.

Zhang, F. and Gao, C. (2020). Convergence rates of empirical Bayes posterior distributions: a variational perspective. `https://arxiv.org/abs/2009.03969v1`.