Power-Expected-Posterior Priors for Variable Selection in Gaussian Linear Models

Dimitris Fouskakis, Ioannis Ntzoufras and David Draper*

*Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu
www.ams.ucsc.edu/∼draper

ISBA 2012
Kyoto, Japan

27 June 2012

Application

• **You** (**Good**, 1950: a **person** wishing to **behave sensibly** in the **presence** of **uncertainty**) have **gathered** a **regression-style data set** (Breiman and Friedman, 1985) to **study** the **relationship** between **daily maximum atmospheric ozone concentration** and a **number** of **meteorological variables** measured the **previous day**, including **temperature**, **wind speed**, **humidity** and **atmospheric pressure**; **data** (**one observation** per **day**: $n = 365$) are from a **variety** of **locations** in the **Los Angeles basin** in **calendar 1976**.

The **outcome variable** w is the **daily maximum** of **24 hourly ozone averages** (**midnight**–1am, **1**–2am, ..., **11pm**–midnight) and has **substantial positive skew**; **defining** $y = \log w$ as the **response variable** to be **modeled** yields **approximate normality**.

Available predictors include **month**, **day of month**, **day of week**, **temperature** (°F) at **Sandburg Air Force Base** and at **El Monte**, **500 mb pressure height** at **Vandenberg Air Force Base**, and **six variables** measured at **Los Angeles International Airport** (LAX): **humidity**, **inversion base height**, **inversion base temperature**, **wind speed**, **visibility** and **pressure gradient** from **LAX** to **Daggett Airport**.

Preliminary Analyses

You perform **descriptive analyses** and **draw** the following **conclusions**:

- **Day of week** has **no effect**; You **combine month** and **day** to create a **day-of-year variable**, which **proxies** for **one form** of **time trend** in a **regression framework**.

- **Temperature at Sandburg** and **temperature at El Monte** are **highly correlated**, and the **El Monte temperature variable** had **139 missing values** (versus only **2 missing values** for the **Sandburg temperature)**, so You **drop** the **El Monte temperature variable**.

- **Omitting all rows of data** for which **one or more** of the **predictors** are **missing** leaves $n = 330$ **days of data** with **no missingness** on the **9 predictor variables** or the **outcome**.

- In **constructing** a **polynomial response surface**, **local-regression** (`loess`) **descriptive analyses** of the **relationships** between **log ozone** and **each** of the **9 predictor variables** reveals **cubic relationships between** the **outcome** and the **predictors** `temp_sandburg` and `inversion_temp`.

Bayesian Variable Selection

- With **9 main effects** there are $\frac{9 \cdot 8}{2} = 36$ **pairwise interactions** among the **main effects**; the **total set** of **potentially useful predictor variables** therefore has **9 main effects, 9 quadratic terms, 2 cubic terms**, and **36 two-way interactions**, for a **total** of **56 predictors**.

Q: **With** the **goal** of **maximally accurate prediction** for **future data** (assumed **conditionally exchangeable** with the **present data, given** the **covariates**), what is the **optimal subset** of these **56 predictors**?

(An **alternative analysis** might involve **Gaussian-process regression modeling**; it would be **interesting** to **compare predictive performance** of the **two approaches** (this is **ongoing work**).)

With an **approximately Gaussian outcome variable (log ozone)**, this **looks like** a **familiar Bayesian variable-selection problem**, but **even here** there's **room** for **potential improvement** over **existing methods**.

Consider two models m_ℓ for $\ell = 0, 1$ with **parameters** $\boldsymbol{\theta}_\ell = (\boldsymbol{\beta}_\ell, \sigma^2_\ell)$ and **likelihood specified** by

$$(\mathbf{Y} | \mathrm{X}_\ell, \boldsymbol{\beta}_\ell, \sigma^2_\ell, m_\ell) \sim N_n(\mathrm{X}_\ell \boldsymbol{\beta}_\ell, \sigma^2_\ell \mathrm{I}_n), \quad (1)$$

Bayesian Model Comparison

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is the **outcome vector**, X_ℓ is an $n \times d_\ell$ **design matrix** containing the **values** of the **predictors** in its **columns**, I_n is the $n \times n$ **identity matrix**, $\boldsymbol{\beta}_\ell$ is a **vector** of length d_ℓ **summarizing** the **effects** of the **covariates** on the **response Y** and σ^2_ℓ is the **error variance** for **model** m_ℓ.

Repeated application of **any method** for **answering** the **question**

Q: Is **model** m_0 **better** than **model** m_1?

will **identify** the **optimal predictor subset** among the 2^p **possibilities** with p **covariates** to **choose from**.

As I've **argued elsewhere** (e.g., Draper 2012, *Adrian Smith Volume*), the **question above begs** a **more fundamental question — better for what purpose? —** and this **turns Bayesian model comparison** into a **decision problem** that **should** be **solved** by **maximizing expected utility**, with a **utility function** that's **closely tailored** to the **specific scientific problem under study**.

However, this is **hard work** (see, e.g., Fouskakis and Draper 2008,

Bayes Factors

JASA); there's a **powerful desire** for **methods** based on **generic utility functions**.

Two such **classes** of **methods** are

{**Bayes factors, BIC**} and {**DIC, log scores**},

grouped by **similarity** of **behavior** in **false-positive** and **false-negative error rates**.

It **turns out** (Draper 2012, submitted) that **neither group uniformly dominates the other** in these **error rates**; **You** have to **choose a method** that's **sensitive** to the **real-world consequences** of **these errors** in **Your problem**.

Here I **focus** on **Bayes factors** and **posterior model probabilities**, **which in turn** require **computation** of **marginal likelihoods**.

When **little information external** to the **present data set** about the **parameter vector** $\theta_\ell = (\beta_\ell, \sigma_\ell^2)$ is **available**, it's **long been known** that **marginal likelihoods** are **hideously sensitive** to the **precise form** in which this **diffuse prior information** is **specified.**

Expected Posterior Priors

Many methods have been **proposed** to **minimize** the **effects** of this **problem**; here I **focus** on an **improvement** that **Fouskakis, Ntzoufras** and **I** have **recently developed** to an **approach** called **expected-posterior priors** (EPPs).

Perez and **Berger** (2002) **defined** the **EPP** as the **posterior distribution** of the **parameter vector** for the **model** under **consideration**, **averaged over all possible imaginary training samples** $\mathbf{y}^*$ **coming** from a **"suitable" predictive distribution** $m^*(\mathbf{y}^*)$.

Thus the **EPP** for the **parameters** of **any model** $m_\ell \in \mathcal{M}$, with $\mathcal{M}$ **denoting** the **model space**, is

$$\begin{aligned} \pi_\ell^E(\boldsymbol{\theta}_\ell) &= \int f(\boldsymbol{\theta}_\ell|\mathbf{y}^*, m_\ell)\, m^*(\mathbf{y}^*)\, d\mathbf{y}^* \\ &= \int \pi_\ell^N(\boldsymbol{\theta}_\ell|\mathbf{y}^*)\, m^*(\mathbf{y}^*)\, d\mathbf{y}^*\,, \end{aligned} \tag{2}$$

where $\pi_\ell^N(\boldsymbol{\theta}_\ell|\mathbf{y}^*)$ is the **posterior** of $\boldsymbol{\theta}_\ell$ for **model** m_ℓ using a **baseline prior** $\pi_\ell^N(\boldsymbol{\theta}_\ell)$ and **"data"** $\mathbf{y}^*$.

EPPs: Choosing the Predictive Distribution

Q: **Which predictive distribution m^* should You use** for the **imaginary data $\mathbf{y}^*$** in (2)?

An **attractive choice**, **leading** to the so-called **base-model approach**, arises from **selecting** a **"reference"** or **"base" model** m_0 for the **training sample** and **defining** $m^*(\mathbf{y}^*) = m_0^N(\mathbf{y}^*) \equiv f(\mathbf{y}^*|m_0)$ to be the **prior predictive distribution**, **evaluated** at $\mathbf{y}^*$, for the **reference model** m_0 under the **baseline prior** $\pi_0^N(\boldsymbol{\theta}_0)$.

Then, for the **reference model** (i.e., when $m_\ell = m_0$), (2) **reduces** to $\pi_0^E(\boldsymbol{\theta}_0) = \pi_0^N(\boldsymbol{\theta}_0)$.

Intuitively the **reference model** should be **at least as simple** as the **other competing models**, and **therefore** a **reasonable choice** is to take m_0 to be a **common sub-model** of all $m_\ell \in \mathcal{M}$.

In the **variable-selection problem** considered **here**, the **constant model** (with **no predictors**) is **clearly a good reference model** that's **nested** in **all the models under consideration**.

This selection makes **calculations simpler**, and **additionally**

makes the **EPP method essentially equivalent** (when n is large) to the **expected intrinsic Bayes factor (EIBF)** approach of **Berger** and **Pericchi** (1996).

One of the **advantages** of **using EPPs** is that there's **no problem** with the **use** of an **improper baseline prior** $\pi_\ell^N(\boldsymbol{\theta}_\ell)$ in (2); the **arbitrary constants cancel out** in the **calculation** of **any Bayes factor**.

Impropriety in m^* also does **not cause indeterminacy**, because m^* is **common** to the **EPPs** for **all models**.

Q: **How** does the **EPP story** go in **variable selection** in **Gaussian regression models**?

$$(\mathbf{Y}|\mathrm{X}_\ell, \boldsymbol{\beta}_\ell, \sigma_\ell^2, m_\ell) \sim N_n(\mathrm{X}_\ell \boldsymbol{\beta}_\ell, \sigma_\ell^2 \mathrm{I}_n) \tag{3}$$

Suppose You have **training data** $\mathbf{y}^*$, of **sample size** n^*, and **design matrix** X^* of **size** $n^* \times p$, where p **denotes** the **total number** of **available covariates**.

Then the **EPP distribution**, given by (2), will **depend** on X^* but **not** on $\mathbf{y}^*$, since the **latter** is **integrated out**.

Difficulties With Training Samples

Q: **How** should You **choose** the **training sample** X^*, and **how big** should it **be**?

The **selection** of a **minimal training sample** has been **proposed**, to **make** the **information content** of the **prior as small as possible**, and **this** is an **appealing idea** on the **surface**.

However, even the **definition** of **"minimal"** turns out to be **open to question**, since it's **problem-specific** (**which models** are You **comparing**?) and **data-specific** (**how many variables** are You **considering**?).

- For **example**, if we **define "minimal"** in **terms** of the **largest model** in **every pairwise comparison**, then the **prior** will **change** in **every comparison**, making the **overall variable-selection procedure incoherent**.

- **Another idea** is to **let** the **size** of the **full model specify** the **minimal training sample**; this **choice** makes **inference** within the **current data set coherent**, but **what happens** if **additional variables** become **available later** in the **study**?

Difficulties With Training Samples (continued)

In **such cases**, the **size** of the **training sample** and hence the **prior** must be **changed**, and the **overall inference** is **again incoherent**.

• **Moreover**, when the **sample size** is **not much larger** than the **number of covariates**, **working** with a **minimal training sample** can **result** in an **unintentionally (highly) influential prior**.

• **Finally**, if the **data derive** from a **highly structured situation**, such as a **complete randomized-blocks experiment**, any **choice** of a **small part** of the **data** to **act** as a **training sample** would be **untypical** of the **full** X **matrix**.

Even if the **minimal-training-sample idea** is **accepted**, the **problem** of **choosing** such a **subset** of the **full data set** still **remains**.

• A **natural solution** involves **computing** the **arithmetic mean** (or some **other summary** of **distributional center**) of the **Bayes factors** over **all possible training samples**, but **this approach** can be **computationally infeasible**, especially when the **number** n of **observations** is **much larger** than the **number** p of **covariates**; for **example**, with $(n, p) = (100, 50)$ and $(500, 100)$ there are **about** 10^{29} and 10^{108} **possible training samples**, respectively.

Difficulties With Training Samples (continued)

• An **obvious choice** at **this point** is to **take** a **random sample** from the **set** of **all possible minimal training samples**, but this **adds** an **extraneous layer** of **Monte-Carlo noise** to the **model-comparison process**.

$\longrightarrow$ The **bottom line**: **EPPs** and **EIBFs get around** the **problems** with **Bayes factors** arising from **diffuse priors**, but **current implementations** of them **suffer** from **difficulties** arising from the **choice** of **training samples**.

Fouskakis, Ntzoufras and **I** have **figured out** how to **solve this problem**, as **follows**.

To **remove** the **sensitivity** of the **EPP method** to the **choice** of **training sample**, we **combine ideas** from the **power-prior approach** of **Ibrahim** and **Chen** (2000) and the **unit-information-prior approach** of **Kass** and **Wasserman** (1995).

As **noted previously**, the **EPP distribution** is the **integral** (over the **training data** $\mathbf{y}^*$) of the **product** of a **posterior distribution** and a **prior predictive distribution**, **both** of **which** have **likelihoods hidden inside them**.

Power-Expected-Posterior (PEP) Priors

As a **first step**, following **Ibrahim** and **Chen**, we **raise both** of these **likelihoods** to the **power** $\frac{1}{\delta}$ and **density-normalize**.

Then, following **Kass** and **Wasserman**, we **set** the **power parameter** δ **equal** to the **training sample size** n^*, to **represent information equal** to **one data point**; in **this way** the **prior corresponds** to a **sample of size one** with the **same sufficient statistics** as the **observed data**.

n^* can be **any integer** from $(p+2)$
(the **minimal training sample size**) to n.

We've **found** that **significant advantages** (and **no disadvantages**) **arise** from the **choice** $n^* = n$, from which $\mathrm{X}^* = \mathrm{X}$: in **this way** we **completely avoid the selection of a training sample** and its **effects** on **posterior model comparison**, while **still holding** the **prior information content** at **one data point**.

In **more detail**, for any $m_\ell \in \mathcal{M}$, we **denote** by $\pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma_\ell^2|\mathrm{X}_\ell^*)$ the **baseline prior** for **model parameters** $\boldsymbol{\beta}_\ell$ and σ_ℓ^2.

Then the **power-expected-posterior** (PEP) prior $\pi_\ell^{PE}(\boldsymbol{\beta}_\ell, \sigma_\ell^2|\,\mathrm{X}_\ell^*\,, \delta)$ **takes** the **following form**:

PEP Priors (continued)

$$\pi_\ell^{PE}(\boldsymbol{\beta}_\ell, \sigma_\ell^2 \,|\, \mathrm{X}_\ell^* \,, \delta) \;=\; \pi_\ell^{N}(\boldsymbol{\beta}_\ell, \sigma_\ell^2 | \mathrm{X}_\ell^*) \int \frac{m_0^N(\mathbf{y}^* | \, \mathrm{X}_0^* \,, \delta)}{m_\ell^N(\mathbf{y}^* | \, \mathrm{X}_\ell^* \,, \delta)} \times f(\mathbf{y}^* | \, \boldsymbol{\beta}_\ell \,, \sigma_\ell^2, m_\ell \,; \mathrm{X}_\ell^* \,, \delta) \, d\mathbf{y}^* \,, \quad (4)$$

where $f(\mathbf{y}^* | \, \boldsymbol{\beta}_\ell \,, \sigma_\ell^2, m_\ell \,; \mathrm{X}_\ell^* \,, \delta) \propto f(\mathbf{y}^* | \boldsymbol{\beta}_\ell \,, \sigma_\ell^2, m_\ell \,; \mathrm{X}_\ell^*)^{\frac{1}{\delta}}$ **is the likelihood raised** to the **power** $\frac{1}{\delta}$ and **density-normalized**, i.e.,

$$\begin{aligned} f(\mathbf{y}^* | \, \boldsymbol{\beta}_\ell \,, \sigma_\ell^2, m_\ell \,; \mathrm{X}_\ell^* \,, \delta) \;&=\; \frac{f(\mathbf{y}^* | \boldsymbol{\beta}_\ell, \sigma_\ell^2, m_\ell \,; \mathrm{X}_\ell^*)^{\frac{1}{\delta}}}{\int f(\mathbf{y}^* | \boldsymbol{\beta}_\ell, \sigma_\ell^2, m_\ell \,; \mathrm{X}_\ell^*)^{\frac{1}{\delta}} \, d\mathbf{y}^*} \\ &=\; \frac{f_{N_{n^*}}(\mathbf{y}^* \,; \, \mathrm{X}_\ell^* \boldsymbol{\beta}_\ell \,, \sigma_\ell^2 \mathrm{I}_{n^*})^{\frac{1}{\delta}}}{\int f_{N_{n^*}}(\mathbf{y}^* \,; \, \mathrm{X}_\ell^* \boldsymbol{\beta}_\ell \,, \sigma_\ell^2 \mathrm{I}_{n^*})^{\frac{1}{\delta}} \, d\mathbf{y}^*} \\ &=\; f_{N_{n^*}}(\mathbf{y}^* \,; \, \mathrm{X}_\ell^* \boldsymbol{\beta}_\ell \,, \delta \, \sigma_\ell^2 \mathrm{I}_{n^*}) \,; \end{aligned} \quad (5)$$

here $f_{N_d}(\mathbf{y} \,; \, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the **density** of the d-**dimensional Normal distribution** with **mean vector** $\boldsymbol{\mu}$ and **covariance matrix** $\boldsymbol{\Sigma}$, evaluated at $\mathbf{y}$.

The **distribution** $m_\ell^N(\mathbf{y}^* | \, \mathrm{X}_\ell^* \,, \delta)$ **appearing** in (4) is the **prior predictive distribution** (or **marginal likelihood**), **evaluated** at $\mathbf{y}^*$, of **model** m_ℓ

PEP Priors (continued)

with the **power likelihood defined** in (5) under **the baseline prior** $\pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma_\ell^2 \,|\, \mathrm{X}_\ell^*)$, i.e.,

$$\begin{aligned} m_\ell^N(\mathbf{y}^* | \, \mathrm{X}_\ell^* \, , \delta) \;\; &= \;\; \int\!\!\int f(\mathbf{y}^* \,|\, \boldsymbol{\beta}_\ell \, , \sigma_\ell^2, m_\ell \, ; \mathrm{X}_\ell^* \, , \delta) \\ &\qquad \times \pi_\ell^N(\boldsymbol{\beta}_\ell \, , \sigma_\ell^2 \,|\, \mathrm{X}_\ell^*) \, d\boldsymbol{\beta}_\ell \, d\sigma_\ell^2 \\ &= \;\; \int\!\!\int f_{N_{n^*}}(\mathbf{y}^* \, ; \, \mathrm{X}_\ell^* \boldsymbol{\beta}_\ell \, , \delta \, \sigma_\ell^2 \mathrm{I}_{n^*}) \\ &\qquad \times \pi_\ell^N(\boldsymbol{\beta}_\ell \, , \sigma_\ell^2 \,|\, \mathrm{X}_\ell^*) \, d\boldsymbol{\beta}_\ell \, d\sigma_\ell^2 \, . \end{aligned} \tag{6}$$

Under the **PEP prior distribution** (4), the **posterior distribution** of the **model parameters** $(\boldsymbol{\beta}_\ell \, , \sigma_\ell^2)$ is

$$\begin{aligned} \pi_\ell^{PE}(\boldsymbol{\beta}_\ell, \sigma_\ell^2 | \mathbf{y}; \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta) \;\; &\propto \;\; f(\mathbf{y} | \boldsymbol{\beta}_\ell \, , \sigma_\ell^2, m_\ell \, ; \mathrm{X}_\ell) \, \pi_\ell^{PE}(\boldsymbol{\beta}_\ell, \sigma_\ell^2 | \mathrm{X}_\ell^*, \delta) \\ &\propto \;\; \int f(\mathbf{y} | \boldsymbol{\beta}_\ell \, , \sigma_\ell^2, m_\ell \, ; \mathrm{X}_\ell) \\ &\qquad \times f(\boldsymbol{\beta}_\ell \, , \sigma_\ell^2 | \mathbf{y}^*, m_\ell \, ; \, \mathrm{X}_\ell^* \, , \delta) \\ &\qquad \times m_0^N(\mathbf{y}^* | \mathrm{X}_0^* \, , \delta) \, d\mathbf{y}^* \, , \end{aligned} \tag{7}$$

and this **last expression equals**

PEP Priors (continued)

$$\int f(\boldsymbol{\beta}_\ell, \sigma^2_\ell | \mathbf{y}, \mathbf{y}^*, m_\ell ; \mathrm{X}_\ell, \mathrm{X}^*_\ell, \delta)\, m^N_\ell(\mathbf{y}|\mathbf{y}^*; \mathrm{X}_\ell, \mathrm{X}^*_\ell, \delta) \times m^N_0(\mathbf{y}^*|\mathrm{X}^*_0, \delta)\, d\mathbf{y}^* , \quad (8)$$

where $f(\boldsymbol{\beta}_\ell, \sigma^2_\ell | \mathbf{y}, \mathbf{y}^*, m_\ell ; \mathrm{X}_\ell, \mathrm{X}^*_\ell, \delta)$ and $m^N_\ell(\mathbf{y}|\mathbf{y}^*; \mathrm{X}_\ell, \mathrm{X}^*_\ell, \delta)$ are the **posterior distribution** of $(\boldsymbol{\beta}_\ell, \sigma^2_\ell)$ and the **marginal likelihood** of **model** m_ℓ, respectively, using **data y** and **design matrix** X_ℓ under **prior** $f(\boldsymbol{\beta}_\ell, \sigma^2_\ell | \mathbf{y}^*, m_\ell ; \mathrm{X}^*_\ell, \delta)$, i.e., the **posterior** of $(\boldsymbol{\beta}_\ell, \sigma^2_\ell)$ with **power Normal likelihood** (5) and **baseline prior** $\pi^N_\ell(\boldsymbol{\beta}_\ell, \sigma^2_\ell | \mathrm{X}^*_\ell)$.

We've **worked out** the **details** (Fouskakis D, Ntzoufras I, Draper D (2012). **Power-expected-posterior priors for variable selection in Gaussian linear models**. Submitted; **copy on my web page**) for the **PEP prior** using **two specific baseline choices**: the **usual independence Jeffreys prior (improper; J-PEP)** and **Zellner's g-prior (proper; Z-PEP)**.

• The **prior mean vector** and **covariance matrix** of $\boldsymbol{\beta}_\ell$, and the **prior mean** and **variance** of σ^2_ℓ, can be **calculated analytically** in the **Z-PEP case**; this **aids** in **prior specification** if **non-diffuse prior information** is **available**.

PEP Priors (continued)

• If **little is known** about the **parameters external** to the **data set**, with **Z-PEP** we **recommend** that the **parameter** g in the **Normal baseline prior** be **set** to $\delta\, n^*$, **so that** with $\delta = n^*$ we **use** $g = (n^*)^2$.

This **choice** will make the g-**prior contribute information** equal to **one data point** within the **posterior** $f(\boldsymbol{\beta}_\ell\,, \sigma^2_\ell | \mathbf{y}^*, m_\ell\,; \mathrm{X}^*_\ell\,, \delta)$.

In **this manner**, the **entire PEP prior accounts** for **information equal** to $\left(1 + \frac{1}{\delta}\right)$ **data points**.

We **suggest setting** the **parameters** a and b in the **Inverse-Gamma baseline prior** to a **small positive value** ϵ, **yielding** a **baseline prior mean** of **1** and **variance** of $\frac{1}{\epsilon}$ (i.e., a **large amount** of **prior uncertainty**) for the **precision parameter**.

• We've **developed simple** and **efficient MCMC algorithms**

(a) to **sample** from the **PEP posterior distributions** and

(b) to **approximate** the **PEP marginal likelihoods needed** to **compute PEP Bayes factors**, in **settings in which** these **marginal likelihoods** are **not analytically tractable**.

PEP Priors (continued)

• **There are** 2^p **models** to **examine** in **finding** the **best subsets** of **predictors**, and **when** p **is large** this **number** is **too big** to **permit exhaustive enumeration** of **marginal likelihoods** for **all subsets**.

When the **marginal likelihood** is **available** in **closed form**, we **use** the **MCMC model composition** (MC^3: Madigan and York (1995)) **method** — a **simple Metropolis algorithm** — to **explore large model spaces**; when **MCMC evaluation** of **marginal likelihoods** is **necessary**, we've **developed** a **Metropolis-within-Gibbs modification** of MC^3 that **exploits** the **structure** of our **MCMC marginal-likelihood estimates**.

• **We have two sets** of **results** that **illustrate PEP in action**:

(1) a **known-truth environment** based on the **simulated data set** of **Nott** and **Kohn** (2005), and

(2) the **LA ozone meteorological application outlined** at the **beginning** of the **talk**.

In **these results** we **compare J-PEP** and **Z-PEP** with the **following previously-developed variable-selection methods**:

Results

(a) the **expected-posterior prior (EPP)** with **minimal training sample**, using the **independence Jeffreys prior** as **baseline** (**call** this **approach J-EPP**) and

(b) **intrinsic Bayes factors (IBFs)** and **expected IBFs (EIBFs)**, i.e., the **arithmetic mean** of **IBFs** over **different minimal training samples**.

Since **Perez** and **Berger** (2002) have **shown** that **Bayes factors** from **J-EPP** become **identical** to **those** from **EIBF** as the **sample size** $n \to \infty$ (with the **number** of **covariates** p **fixed**), it's **possible** (for large n) to **use EIBF** as an **approximation** to **J-EPP** that's **computationally much faster** than the **full J-EPP calculation**.

For this **reason, You** can **regard** the **labels "J-EPP"** and **"EIBF"** as **more or less interchangeable** in **what follows**.

Nott-Kohn example: This **data set consists** of $n = 50$ **observations** with $p = 15$ **covariates**.

The **first 10 covariates** are **generated** from a **multivariate Normal distribution** with **mean vector 0** and **covariance matrix** I_{10}, while

Results: Nott-Kohn Data Set

$$X_{ij} \sim N\left(0.3X_{i1} + 0.5X_{i2} + 0.7X_{i3} + 0.9X_{i4} + 1.1X_{i5}, 1\right) \quad (9)$$

for $(j = 11, \ldots, 15; i = 1, \ldots, 50)$, and the **response** is **generated** from

$$Y_i \sim N\left(4 + 2X_{i1} - X_{i5} + 1.5X_{i7} + X_{i,11} + 0.5X_{i,13}, 2.5^2\right) \quad (10)$$

for $i = 1, \ldots, 50$.

With $p = 15$ **covariates** there are **only 32,768 models** to **compare**; a **full enumeration** of the **model space** is **possible** (MC^3 is **not needed**).

The **tables** on the **next page** present **posterior model probabilities** for the **best models** and **posterior variable-inclusion probabilities**, together with **Bayes factors** of the **MAP model** (m_1) **against** m_j $(j = 2, \ldots, 7)$, for **Z-PEP** and **J-PEP**.

The **data-generating model** M_{DG} is $(X_1 + X_5 + X_7 + X_{11} + X_{13})$, but **this example** is **tricky because** X_{13} has a **much smaller effect** than (X_1, X_5, X_7, X_{11}).

Z-PEP and **J-PEP** both **fail to identify** M_{DG} as the **MAP model** (the **same** is **true** of **J-EPP** and **EIBF**), but **all** of their **high-posterior-probability models** are **reasonable**; **J-PEP** is **somewhat more parsimonious** than **Z-PEP**.

Results: Nott-Kohn Data Set

Posterior model probabilities for the best models, together with Bayes factors of the MAP model (m_1) against m_j ($j = 2, \ldots, 7$), for the Z-PEP and the J-PEP prior methodologies in the simulated example of Nott and Kohn.

		Z-PEP		J-PEP		
m_j	Predictors	Posterior Model Probability	Bayes Factor	Rank	Posterior Model Probability	Bayes Factor
1	$X_1 + X_5 + X_7 + X_{11}$	0.0783	1.00	(2)	0.0952	1.00
2	$X_1 + X_7 + X_{11}$	0.0636	1.23	(1)	0.1054	0.90
3	$X_1 + X_5 + X_6 + X_7 + X_{11}$	0.0595	1.32	(3)	0.0505	1.88
4	$X_1 + X_6 + X_7 + X_{11}$	0.0242	3.23	(4)	0.0308	3.09
5	$X_1 + X_7 + X_{10} + X_{11}$	0.0175	4.46	(5)	0.0227	4.19
6	$X_1 + X_5 + X_7 + X_{10} + X_{11}$	0.0170	4.60	(9)	0.0146	6.53
7	$X_1 + X_5 + X_7 + X_{11} + X_{13}$	0.0163	4.78	(10)	0.0139	6.87

Posterior variable-inclusion probabilities for the Z-PEP and J-PEP prior methodologies for the simulated example of Nott and Kohn.

	Covariate							
Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Z-PEP	0.997	0.110	0.129	0.133	0.503	0.337	1.000	0.150
J-PEP	0.993	0.088	0.108	0.121	0.395	0.253	1.000	0.117

	Covariate						
Method	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
Z-PEP	0.126	0.197	0.856	0.136	0.142	0.111	0.113
J-PEP	0.100	0.152	0.789	0.109	0.115	0.099	0.100

Sensitivity Analysis for the Training Sample Size n^*

To **examine** the **sensitivity** of the **PEP approach** to the **sample size** n^* of the **training data set**, we **present results** for $n^* = 17, \ldots, 50$.

The **figures** on the **next two pages** display **posterior marginal variable-inclusion probabilities** and **posterior model probabilities, respectively**.

As **noted previously**, to **specify** X^* when $n^* < n$ we **randomly selected** a **sub-sample** of the **rows** of the **original matrix** X.

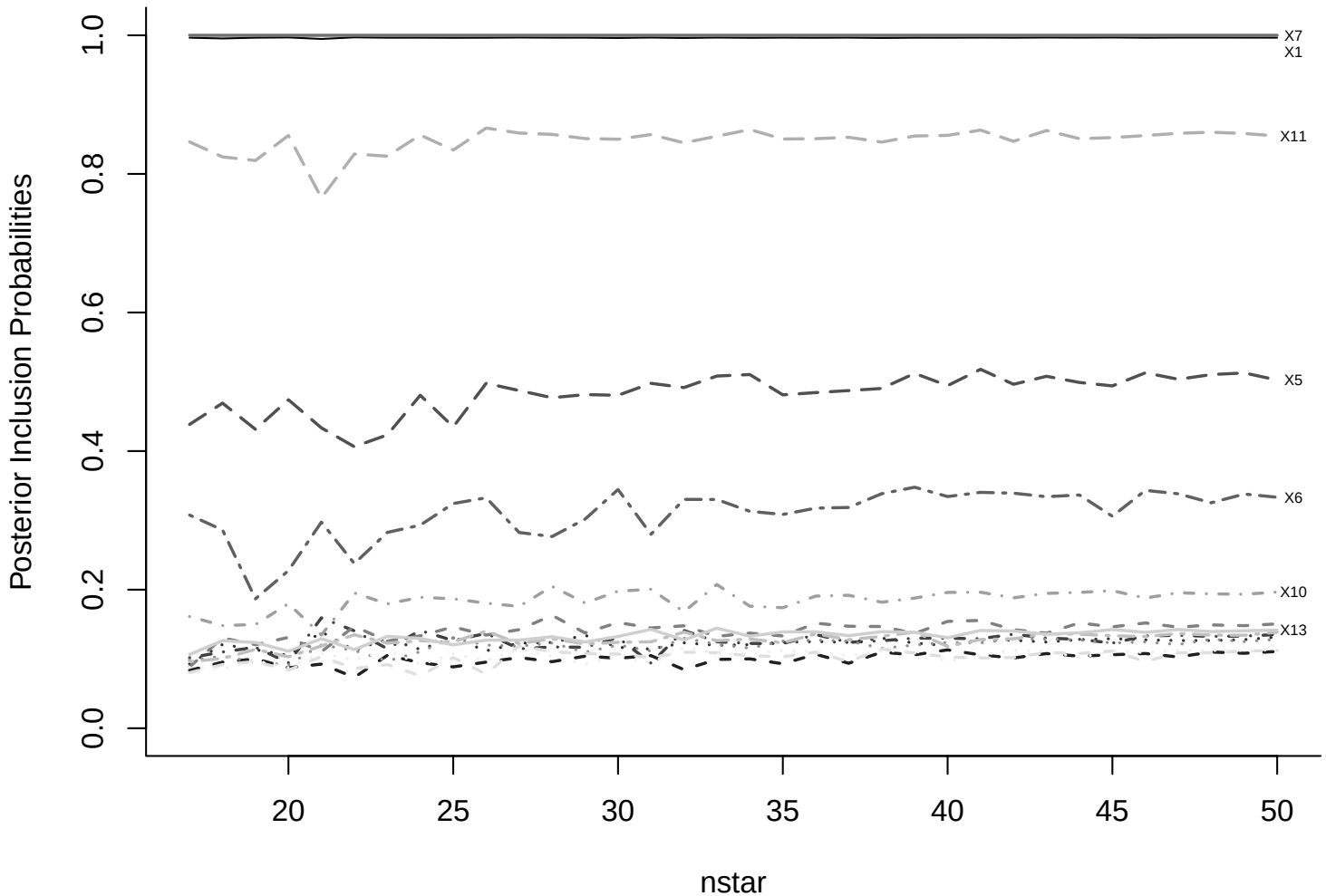
Results are **presented** for **Z-PEP**; **J-PEP** produced **similar findings**:

• Both **posterior inclusion probabilities** and**posterior model probabilities** are **quite insensitive** to a **wide variety of values** of n^*.

• **Therefore** we can **use** $n^* = n$ and **dispense with training samples altogether**; this **yields** the following **advantages**: **increased stability** of the **resulting Bayes factors**, **removal** of the **arbitrariness arising** from **individual training-sample selections**, and **substantial increases** in **computational speed**, allowing **many more models** to be **compared** within a **fixed CPU budget**.

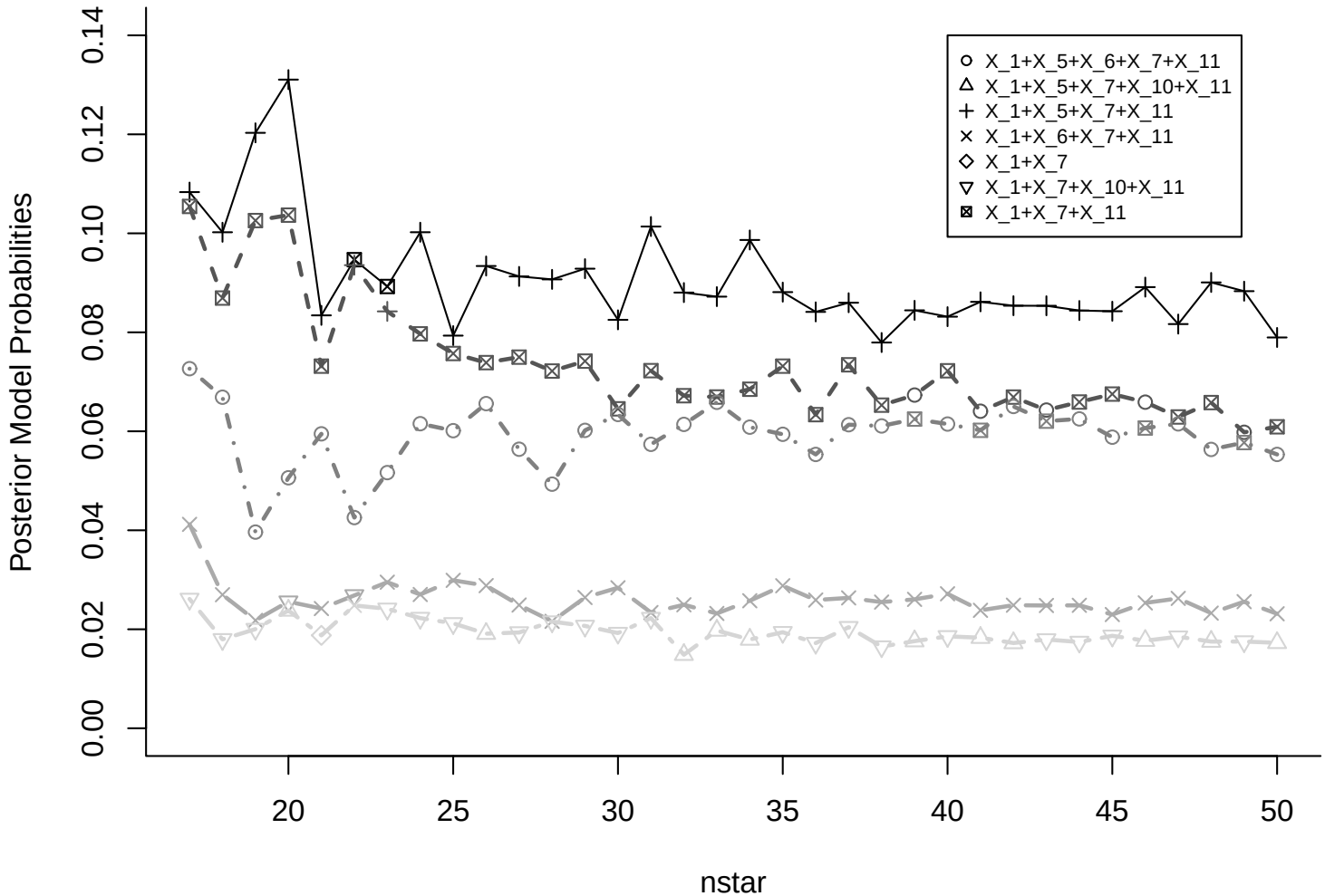
Sensitivity Analysis for the Training Sample Size n^*

Posterior marginal inclusion probabilities for different n^* with the Z-PEP prior methodology for the simulated example of Nott and Kohn.



Sensitivity Analysis for the Training Sample Size n^*

Posterior model probabilities of the five best models obtained for each n^*, with the Z-PEP prior methodology for the simulated example of Nott and Kohn.



Comparisons With IBF and J-EPP

Here we **compare** the **PEP Bayes factor between** the **two best models** $((X_1 + X_5 + X_7 + X_{11})$ and $(X_1 + X_7 + X_{11}))$ with the **corresponding Bayes factors** using **J-EPP** and **IBF**.

For **IBF** and **J-EPP** we **randomly selected 100 training samples** of **size** $n^* = 6$ (the **minimal training sample size** for **comparison** of **these two models**) and $n^* = 17$ (the **minimal training sample size** for the **estimation** of the **full model** with **all** $p = 15$ **covariates**), while for **Z-PEP** and **J-PEP** we **randomly selected 100 training samples** of **sizes** $n^* = 6, 17$ and $n^* = 5 \times k$ for $k = 5, \ldots, 10$.

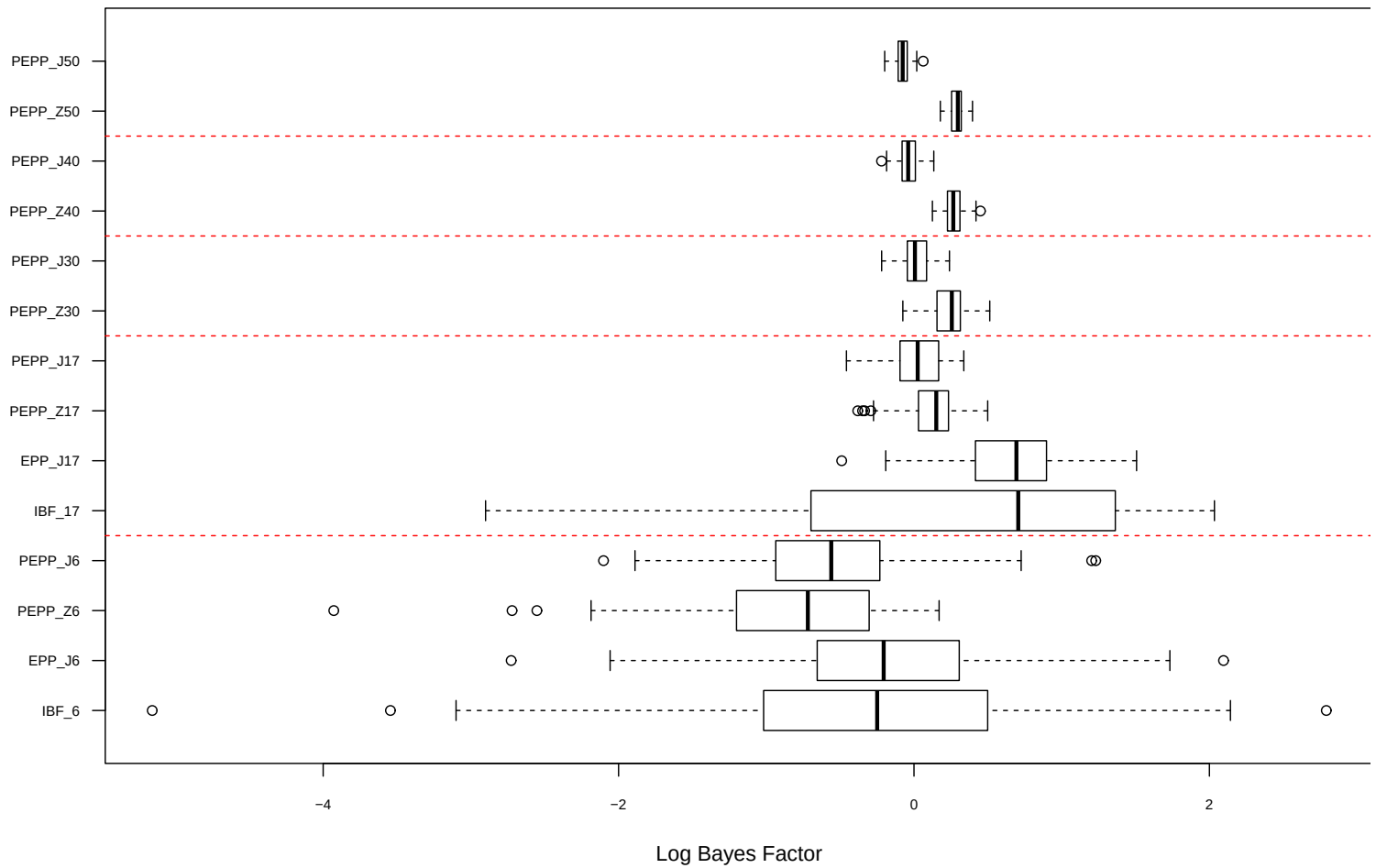
The **figure** on the **next page presents** the **results** as **parallel boxplots**, and **shows** that

- With **modest** n^* **values**, which would **tend** to be **favored** by **users** for their **advantage in computing speed**, the **IBF method exhibited** an **extraordinary amount of instability** across the **particular random training samples chosen**, and the **J-EPP method** was **not much better**; and

- **In contrast, J-PEP** and **Z-PEP** are **highly stable** as a **function** of **training sample**.

Comparisons With IBF and J-EPP (continued)

Boxplots of the Intrinsic Bayes Factor (IBF) and Bayes factors using the J-EPP, J-PEP and Z-PEP approaches, on a logarithmic scale, in favor of model $(X_1 + X_5 + X_7 + X_{11})$ *over model* $(X_1 + X_7 + X_{11})$ *for the simulated example of Nott and Kohn.*



Ozone Application

Full-enumeration search for the **full space** with **56 covariates** was **computationally infeasible**, so we **used** the **model search algorithm** (*MC*^{3}) **described above** for **Z-PEP** and **EIBF**.

With such a large number of predictors, the **model space in our problem** is **too large** for the *MC*^{3} **approach** to **estimate posterior model probabilities** with **high accuracy** in a **reasonable amount of CPU time**.

For this **reason**, we **implemented** the following **two-step method**:

(1) **First we used** *MC*^{3} to **identify variables** with **high posterior marginal inclusion probabilities**, and we **created** a **reduced model space** consisting only of **those variables** whose **marginal probabilities** were **above a threshold value of 0.3**.

(2) **Then** we **used** the **same model search algorithm** as in **step (1)** in the **reduced space** to **estimate posterior model probabilities**.

The **tables** on the **next two pages show** that the **PEP methodology supports more parsimonious models** than the **EIBF approach**.

Ozone Application (continued)

Posterior inclusion probabilities using Z-PEP, J-PEP and EIBF for the reduced model space of the ozone data set.

Index	Name	J-PEP	Z-PEP	EIBF
1	Day of year	1.000	1.000	1.000
2	Wind speed at LAX	0.985	0.992	0.976
5	Temperature at Sandburg	0.182	0.375	0.475
7	PG from LAX to Daggett	0.613	0.857	0.984
8	Inversion base temperature at LAX	1.000	1.000	1.000
9	Visibility at LAX	1.000	1.000	1.000
10	(Day of year)^2	1.000	1.000	1.000
12	(500 mb pressure height at VAFB)^2	0.618	0.840	0.980
13	(Humidity at LAX)^2	0.716	0.918	1.000
15	(Inversion base height at LAX)^2	0.983	0.988	1.000
16	(PG from LAX to Daggett)^2	1.000	1.000	1.000
18	(Visibility at LAX)^2	0.896	0.965	0.995
20	(Inversion base temperature at LAX)^3	0.401	0.641	0.923
23	(Day of year) $\times$ (Humidity at LAX)	0.006	0.011	0.027
26	(Day of year) $\times$ (PG from LAX to Daggett)	0.042	0.093	0.233
30	(Wind speed at LAX) $\times$ (Humidity at LAX)	0.011	0.027	0.073
36	(500 mb pressure height at VAFB) $\times$ (Humidity at LAX)	0.036	0.087	0.100
39	(500 mb pressure height at VAFB) $\times$ (PG from LAX to Daggett)	0.040	0.159	0.371
42	(Humidity at LAX) $\times$ (Temperature at Sandburg)	0.315	0.429	0.694
43	(Humidity at LAX) $\times$ (Inversion base height at LAX)	0.974	0.898	0.877
48	(Temperature at Sandburg) $\times$ (PG from LAX to Daggett)	0.017	0.026	0.028
51	(Inversion base height at LAX) $\times$ (PG from LAX to Daggett)	0.065	0.197	0.342

Ozone Application (continued)

Posterior odds (PO_{1k}) *of the five best models within each analysis versus the current model* k*, for the reduced model space of the ozone data set.*

J-PEP

Ranking			Additional Variables	Number of	Posterior
J-PEP	Z-PEP	EIBF		Covariates	Odds PO_{1k}
1	(>5)	(>5)		9	1.00
2	(1)	(5)	$X_7 + X_{12} + X_{13} + X_{20}$	13	1.29
3	(>5)	(>5)	$X_7 + X_{13} + X_{20}$	12	1.46
4	(>5)	(>5)	$X_{12} + X_{20}$	11	1.87
5	(>5)	(>5)	X_{12}	10	2.08

Z-PEP

Ranking			Additional Variables	Number of	Posterior
Z-PEP	J-PEP	EIBF		Covariates	Odds PO_{1k}
1	(2)	(5)	$X_7 + X_{12} + X_{13} + X_{20}$	13	1.00
2	(>5)	(>5)	$X_5 + X_7 + X_{12} + X_{13} + X_{20}$	14	1.19
3	(>5)	(3)	$X_5 + X_7 + X_{12} + X_{13} + X_{20} + X_{42}$	15	1.77
4	(>5)	(1)	$X_7 + X_{12} + X_{13} + X_{20} + X_{42}$	14	1.94
5	(>5)	(>5)	$X_7 + X_{12} + X_{13}$	12	2.30

EIBF

Ranking			Additional Variables	Number of	Posterior
EIBF	J-PEP	Z-PEP		Covariates	Odds PO_{1k}
1	(>5)	(4)	$X_7 + X_{12} + X_{13} + X_{20} + X_{42}$	14	1.00
2	(>5)	(>5)	$X_5 + X_7 + X_{12} + X_{13} + X_{20} + X_{26} + X_{42}$	16	1.17
3	(>5)	(3)	$X_5 + X_7 + X_{12} + X_{13} + X_{20} + X_{42}$	15	1.30
4	(>5)	(>5)	$X_7 + X_{12} + X_{13} + X_{20} + X_{39} + X_{42}$	15	1.44
5	(2)	(1)	$X_7 + X_{12} + X_{13} + X_{20}$	13	1.58

Ozone Application (continued)

Comparison of the predictive performance of the PEP and J-EPP methods, using the full and MAP models in the reduced model space of the ozone data set.

Model	d_ℓ	R^2	R^2_{adj}	$RMSE^*$ J-PEP	Z-PEP	J-EPP	Jeffreys Prior
Full	22	0.8500	0.8392	0.5988 (0.0087)	0.5935 (0.0097)	0.6194 (0.0169)	0.5972 (0.0104)
J-PEP MAP	9	0.8070	0.8016	0.5975 (0.0063)	0.6161 (0.0051)	0.7524 (0.0626)	0.6165 (0.0052)
Z-PEP MAP	13	0.8370	0.8303	0.5994 (0.0071)	0.5999 (0.0060)	0.6982 (0.0734)	0.5994 (0.0049)
EIBF MAP	14	0.8398	0.8326	0.6182 (0.0066)	0.5961 (0.0072)	0.6726 (0.0800)	0.5958 (0.0061)

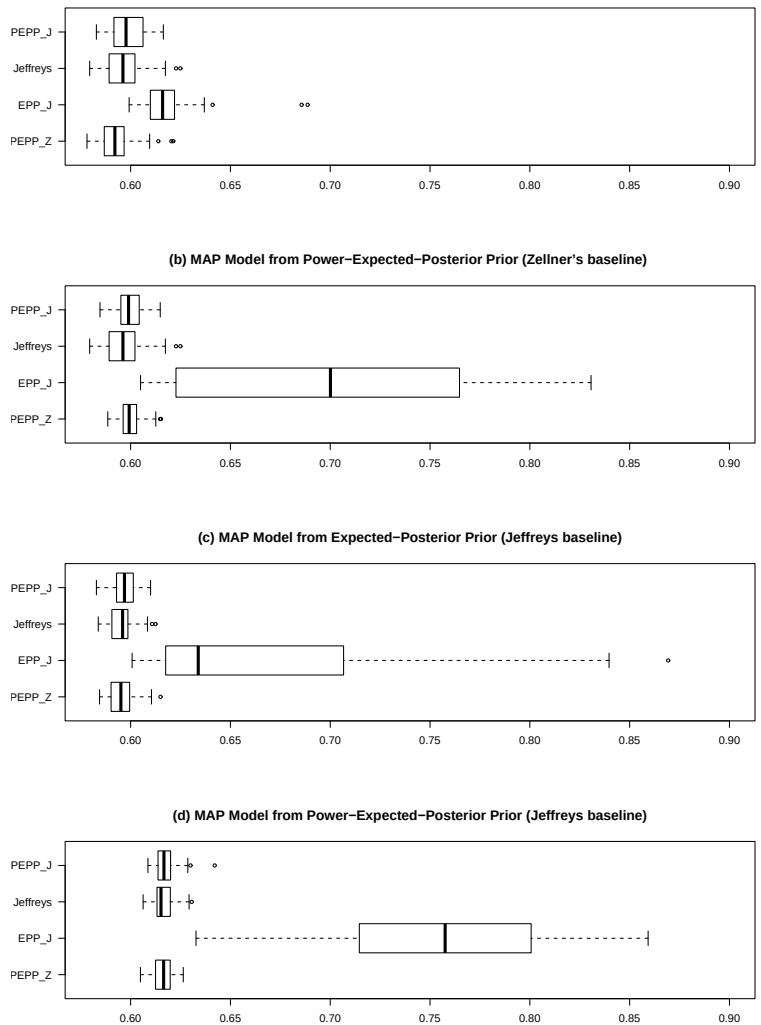
Comparison with the full model (percentage changes)

Model	d_ℓ	R^2	R^2_{adj}	$RMSE$ J-PEP	Z-PEP	J-EPP	Jeffreys Prior
J-PEP MAP	−59%	−5.06%	−4.48%	−0.22%	+3.81%	+21.5%	+3.23%
Z-PEP MAP	−41%	−1.50%	−1.06%	+0.10%	+1.01%	+12.7%	+0.37%
EIBF MAP	−36%	−1.20%	−0.78%	+3.24%	+0.44%	+10.9%	−0.23%

The **PEP priors choose more parsimonious models** than **EIBF** and **J-EPP**, while **simultaneously having better out-of-sample predictive accuracy**; in **fact** they **achieve** the **same predictive accuracy** as the **independence Jeffreys prior (IJP)** applied to the **same models** (and **IJP cannot be used for variable selection** because the **Jeffreys prior is improper**).

Ozone Application (continued)

Distribution of RMSE across 50 random partitions of the ozone data set, for the Jeffreys-prior, J-EPP, Z-PEP and J-PEP methods, in four different models.



Conclusions

The **major contribution** of the **research presented here** is to **sharply diminish the effect of training samples** on **previously-studied expected-posterior-prior and intrinsic-Bayes-factor methodologies**; our method

- is **systematically more parsimonious** (under **either baseline prior choice**) than the **EPP approach** using the **Jeffreys prior** as a **baseline prior** and **minimal training samples**, while **sacrificing no desirable performance characteristics** to **achieve this parsimony**;
- is **robust to the size of the training sample**, thus **supporting the use** of the **entire data set as a "training sample"** and thereby

— **promoting stability** of the **resulting Bayes factors**,

—**avoiding arbitrariness** arising from **individual training-sample selections**, and

— **achieving fast computation**, allowing **many more models to be examined** in a **fixed CPU budget**; and

- **identifies maximum a-posteriori models** that **achieve good out-of-sample predictive performance**.