Model Uncertainty: Why It Matters, and What To Do About It

David Draper

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

draper@ams.ucsc.edu

www.ams.ucsc.edu/∼draper

Avon Local RSS Meeting on Complexity and Uncertainty

10 May 2007

Model Uncertainty

In a typical application of the **statistical paradigm**, there's some quantity Q about which I'm at least partially **uncertain**, and I wish to **quantify** my uncertainty about Q, for the purpose of (a) sharing this information with other people (**inference**) or (b) helping myself or others to make a **choice** in the face of this uncertainty (**decision-making**).

Uncertainty quantification is usually based on a **probability model** M, which relates Q to **known quantities** (such as **data values** D); M will in turn be based on **assumptions** and **judgments** on my part about how Q and D are **related**, but I'm not always certain about the **"right"** assumptions and judgments to make.

To be completely **honest**, then, I have to acknowledge **two sources of uncertainty**: I'm uncertain about Q, and I'm also uncertain about **how to quantify my uncertainty** about Q.

This second source is **model uncertainty**.

Probability

Earlier I mentioned **probability**; what do I **mean** by that?

Two main **probability paradigms**: **frequentist** and **Bayesian**.

• **Frequentist** probability: Restrict attention to phenomena that are **inherently repeatable** under (essentially) **identical conditions**; then, for an event A of interest, $P_F(A) =$ limiting **relative frequency** with which A occurs in the (hypothetical) repetitions, as number of repetitions $n \to \infty$.

• **Bayesian** probability: numerical **weight of evidence** $P_B(A)$ in favor of an uncertain proposition A, obeying a series of **reasonable axioms** to ensure that Bayesian probabilities are **coherent** **(internally logically consistent)**.

People have acted for a long time as if you have to **choose** one of these paradigms and **defend** it against **attacks** from those who prefer the other one, but (a) **both paradigms** have **strengths** and **weaknesses** and (b) nothing **forces** us to **make** such a **choice**: I believe that my job as a statistician is to develop a **fusion** of the two paradigms that **maximizes** the **strengths**, and **minimizes** the **weaknesses**, of the **two approaches**.

Probabilistic Fusion

The **main frequentist strength** is in paying attention to **calibration** (how often do I get the **right answer**?); the main **Bayesian weakness** is that **good calibration** is **not guaranteed** with the Bayesian approach.

This motivates for me the following **fusion**:

(a) **Reason in a Bayesian way** when **formulating** my **inferences** and **decisions** (because the **Bayesian paradigm** is the **most flexible approach** so far invented for **quantifying all relevant sources of uncertainty** in **complicated problems** and **making choices** in the face of such uncertainty);

(b) **Reason in a frequentist way** when **evaluating** the **quality** of these **inferences** and **decisions**, by keeping track of the **calibration** of my **Bayesian methods** (e.g., by constructing **predictive distributions** for **observables** and **comparing** those distributions with the **actual observed values**).

In this way **Bayesian coherence** keeps me **internally free of logical contradiction** and **frequentist calibration** keeps me **honest** (in **good touch** with the **external world**).

What Is a Bayesian Model?

Definition: A **Bayesian model** is a **mathematical framework** (embodying **assumptions** and **judgments**) for **quantifying uncertainty about unknown quantities** by relating them to **known quantities**.

Desirable for the **assumptions** and **judgments** in the model to arise as directly as possible from **contextual information** in the problem under study.

The most satisfying approach to **achieving this goal** appears to be that of de Finetti (1930): a **Bayesian model** is a **joint predictive distribution**

$$p(y) = p(y_1, \ldots, y_n) \tag{1}$$

for as-yet-unobserved **observables** $y = (y_1, \ldots, y_n)$.

Example 1: Data = **health outcomes** for all patients at one hospital with heart attack admission diagnosis.

Simplest possible: $y_i = 1$ if patient i **dies within 30 days of admission**, 0 otherwise.

Exchangeability

de Finetti (1930): **in absence of any other information**, my predictive uncertainty about y_i is **exchangeable** (**symmetric** under **permutation** of the **order** of the observations).

Representation theorem for binary data: if I'm willing to regard $(y_1, \ldots, y_n)$ as part of an **infinitely exchangeable sequence** (meaning that I judge all **finite subsets** exchangeable; this is like **thinking** of the y_i as having been **randomly sampled** from the **population** $(y_1, y_2, \ldots)$), then to be **coherent** my joint predictive distribution $p(y_1, \ldots, y_n)$ must have the simple **hierarchical** form

$$\begin{aligned} \theta &\sim p(\theta) \\ (y_i|\theta) &\overset{\text{IID}}{\sim} \text{Bernoulli}(\theta), \end{aligned} \tag{2}$$

where $\theta = P(y_i = 1) =$ **limiting value of mean of** y_i in infinite sequence.

Mathematically $p(\theta)$ is **mixing distribution** in

$$p(y_1, \ldots, y_n) = \int_0^1 \prod_{i=1}^{n} p(y_i|\theta)\, p(\theta)\, d\theta \ . \tag{3}$$

Model = Exchangeability + Prior

Statistically, $p(\theta)$ provides opportunity to quantify **prior information** about θ and combine with information in y.

Thus, in simplest situation, **Bayesian model specification** = **exchangeability judgment** + choice of **scientifically appropriate prior distribution** $p(\theta)$.

Example 2 (elaborating Example 1): Now I want to predict real-valued **sickness-at-admission score** instead of mortality (still no **covariates**).

Uncertainty about y_i still **exchangeable**; de Finetti's (1937) **representation theorem** for real-valued data: if $(y_1, \ldots, y_n)$ part of **infinitely exchangeable sequence**, all **coherent** joint predictive distributions $p(y_1, \ldots, y_n)$ must have hierarchical form

$$\begin{aligned} F &\sim p(F) \\ (y_i|F) &\overset{\text{IID}}{\sim} F, \end{aligned} \tag{4}$$

where F = **limiting empirical cumulative distribution function** (CDF) of infinite sequence $(y_1, y_2, \ldots)$.

Bayesian Nonparametrics

Thus here **Bayesian model specification = exchangeability judgment +** choice of **scientifically appropriate mixing (prior) distribution** $p(F)$ for F.

However, F is **infinite-dimensional parameter**; putting probability distribution on $\mathcal{D} = \{\text{all possible CDFs}\}$ is harder.

Specifying distributions on **function spaces** is task of **Bayesian nonparametric** (BNP) modeling (e.g., Dey et al. 1998).

Example 3 (elaborating Example 2): In practice, in addition to **outcomes** y_i, **covariates** x_{ij} will typically be available.

For instance (Hendriksen et al. 1984), 572 elderly people **randomized**, 287 to **control** (C) group (standard care) and 285 to **treatment** (T) group (standard care plus **in-home geriatric assessment** (IHGA): **preventive medicine** in which each person's medical/social needs assessed, acted upon individually).

One important **outcome** was **number of hospitalizations** (in two years):
y_i^T, y_j^C = numbers of hospitalizations for **treatment** person i, **control** person j, respectively.

Conditional Exchangeability

Suppose **treatment/control** (T/C) status is **only available covariate**.

Unconditional judgment of exchangeability across all 572 outcomes **no longer automatically scientifically appropriate**.

Instead **design of experiment** compels (at least initially) judgment of **conditional exchangeability given T/C status** (e.g., de Finetti 1938, Draper et al. 1993), as in

$$\begin{array}{ccc} (F_T, F_C) & \sim & p(F_T, F_C) \\ (y_i^T | F_T, F_C) & \overset{\text{IID}}{\sim} & F_T \;\Big|\; (y_j^C | F_T, F_C) \quad \overset{\text{IID}}{\sim} \quad F_C \end{array} \tag{5}$$

This framework, in which (a) **covariates** specify **conditional exchangeability judgments**, (b) de Finetti's **representation theorem** reduces model specification task to placing appropriate prior distributions on CDFs, covers much of field of **statistical inference/prediction**.

Note that even in this **rather general nonparametric framework** it will be necessary to have a **good tool** for **discriminating between the quality of two models**:

Data-Analytic Model Specification

unconditional exchangeability ($F_T = F_C$; T has **same effect** as C) versus **conditional** exchangeability ($F_T \neq F_C$; T and C effects **differ**); the **model discrimination tool** I favor is based on **predictive log scores** (Draper and Krnjajić 2007).

Basic problem of Bayesian **model choice**: Given future observables $y = (y_1, \ldots, y_n)$, I'm **uncertain** about y (**first-order**), but I'm also **uncertain about how to specify my uncertainty** about y (**second-order**); I want to cope with **both of these kinds of uncertainty** in a **well-calibrated** manner.

Standard (**data-analytic**) approach to model specification involves initial choice, for **structure** of model, of **standard parametric family**, followed by **modification** of initial choice—once data begin to arrive—if data suggest **deficiencies** in original specification.

This approach (e.g., Draper 1995) is **incoherent** (unless I pay an **appropriate price** for **shopping around** for the model).

Cromwell's Rule

The **data-analytic** approach uses data both to specify **prior distribution on structure space** and to **update** using **data-determined prior** (result will typically be **uncalibrated** (too narrow) predictive distributions for future data).

Dilemma is example of **Cromwell's Rule** (if $p(\theta) = 0$ then $p(\theta|y) = 0$ for all y): initial model choice placed **0 prior probability** on **large regions of model space**; formally all such regions **must also have 0 posterior probability** even if data indicate **different prior on model space** would have been better.

Three possible solutions:

- **3CV** (a modification of the usual **cross-validation** approach, which solves the problem by **paying an appropriate price for model exploration**),
- **BNP** (which solves the problem by **"not putting zero probability on anything"**), and
- **BMA** (**Bayesian model averaging**, which solves the problem by **averaging over model uncertainty** instead of **ignoring it**; BMA can be thought of as a **parametric approximation** to BNP).

3CV

• **Three-way cross-validation** (3CV; Draper and Krnjajić 2007): taking usual cross-validation idea one step further,

(1) Partition data at random into *three* (non-overlapping and exhaustive) subsets S_i.

(2) Fit tentative {likelihood + prior} to S_1. **Expand** initial model in all feasible ways suggested by data exploration using S_1. **Iterate** until you're happy.

(3) Use final model (fit to S_1) from (2) to create predictive distributions for all data points in S_2. Compare actual outcomes with these distributions, checking for **predictive calibration**. Go back to (2), change likelihood as necessary, **retune prior** as necessary, to get good calibration. **Iterate** until you're happy.

(4) Announce **final model** (fit to $S_1 \cup S_2$) from (3), and report **predictive calibration** of this model on data points in S_3 as indication of how well it would perform with new data.

3CV (continued)

With **large** n probably only need to do this **once**; with **small** and **moderate** n probably best to **repeat** (1–4) several times and **combine** results via **BMA**).

Q: **How large** should the S_i be?

A: Theory suggests a **recommended allocation of data** across (S_1, S_2, S_3) in the vicinity of **(50%, 25%, 25%)**.

In other words, with $n =$ **1,000** you should be prepared to **pay about 250 observations worth of information** in **quoting your final uncertainty assessments** (i.e., making these uncertainty bands **about** $\sqrt{\frac{n}{0.75n}} \doteq 15\%$ **wider** than those based on the full data set, a rule that applies with any n), to **account in a well-calibrated manner** for your **search for a good model**.

NB **All three solutions** to the problem of **accounting for model uncertainty** in a **well-calibrated manner** (3CV, BNP, BMA) will typically lead to **wider uncertainty bands** than those obtained by **pretending you know the "right" model when you don't**; the main difference will be in **how** the bands widen (3CV: explicitly **set aside** some data for **calibration** purposes).

BNP

Simplest **Bayesian nonparametric** model: for $i = 1, \ldots, n$,

$$\begin{aligned} F &\sim p(F) \\ (y_i|F) &\overset{\text{IID}}{\sim} F. \end{aligned} \tag{6}$$

How place a **scientifically relevant prior distribution** on the **underlying CDF** F?

In all Bayesian work, specifying a **prior** for an **unknown quantity** (like F) comes down to choosing **two ingredients** (based on scientific context): a **prior estimate** F_0 for the unknown, and a **prior sample size** n_0 (the prior can typically be thought of as **equivalent to a data set** with some number n_0 of observations).

Two main BNP methods to date:

- **Dirichlet process** (DP) priors, and **Dirichlet process mixture modeling** (DPMM): the DP prior was defined so that it's **conjugate** (the **prior** and **posterior** have the **same form**) and it involves the **empirical CDF** $\hat{F}_n$ (which places mass $\frac{1}{n}$ on each of the data points y_i) in a **natural** way:

BNP (continued)

$$F \sim DP(n_0 F_0),\ (y_i|F) \overset{\text{IID}}{\sim} F \ \rightarrow\ (F|Y) \sim DP(n^* F^*)$$
$$n^* = n_0 + n,\quad F^* = \frac{n_0 F_0 + n\hat{F}_n}{n_0 + n}. \tag{7}$$

It turns out (Ferguson 1973) that **DP priors** place **all their mass** on **discrete distributions**, so — in order to model **continuous outcomes** and permit the prior estimate F_0 to be an entire **parametric family of distributions** — it's natural to work with **DP mixture models** that involve **latent mixing parameters** θ_i, one for each observation:

$$\begin{aligned}
(y_i|\theta_i, \phi) &\overset{\text{indep}}{\sim} p(\cdot; \theta_i, \phi), \quad i = 1, \ldots, n \\
(\theta_i|G) &\overset{\text{IID}}{\sim} G \\
(G|n_0, \psi) &\sim DP(n_0 G_0), G_0 = G_0(\cdot|\psi) \\
(\phi, n_0, \psi) &\sim p(\phi)\, p(n_0)\, p(\psi)
\end{aligned} \tag{8}$$

For example, $p(\cdot; \theta_i, \phi) = N(\theta_i, \phi)$ specifies the **location normal Dirichlet process mixture model** (LNDPMM).

BNP (continued)

• **Pólya trees**: This is another **conjugate** class of **BNP priors** on CDFs, based on the idea of **partitioning** the real line into finer and finer **grids (histogram bars)** and letting the amount of **probability** in each histogram bar itself be **random**.

In the **specification**

$$F \sim PT(\Pi, \mathcal{A}_{n_0}),$$

Π (a **binary tree partition**) plays the role of the **prior estimate** F_0 of F and $\mathcal{A}_{n_0}$ (a collection of **prior weights** on the **random histogram bars**) plays the role of the **prior sample size** n_0.

DPMMs and **Pólya tree models** can be fit via **Markov chain Monte Carlo** methods (MCMC; e.g., Dey, Müller and Sinha 1998), and there are **many extensions** to **more realistically complicated settings** (e.g., **regression, spatial processes, time series modeling**, and **categorical data analysis**).

BNP (continued)

The **Bayesian paradigm** is based on the idea that

posterior information = **prior information** + **data information** ,

i.e., **stronger prior assumptions** → **narrower uncertainty bands**.

Bayesian parametric models are a **special case** of **BNP models** in which (in effect) **more informative priors** on the **space of all CDFs** are **assumed**.

When this **extra prior information** is **"correct,"** the resulting **parametric uncertainty bands** will be **appropriately narrower** than BNP bands; the problem is that when it's **incorrect**, the parametric uncertainty bands will **still be narrower**, but in that case the **extra narrowness** is **inappropriate** and will typically lead to **poor calibration of the parametric models**.

Thus **BNP solves the model uncertainty problem** by making **weaker (and thus likelier to be "correct") prior assumptions** than parametric modeling — as with 3CV, this again **widens the uncertainty bands** in pursuit of **better calibration**.

Bayesian Model Averaging

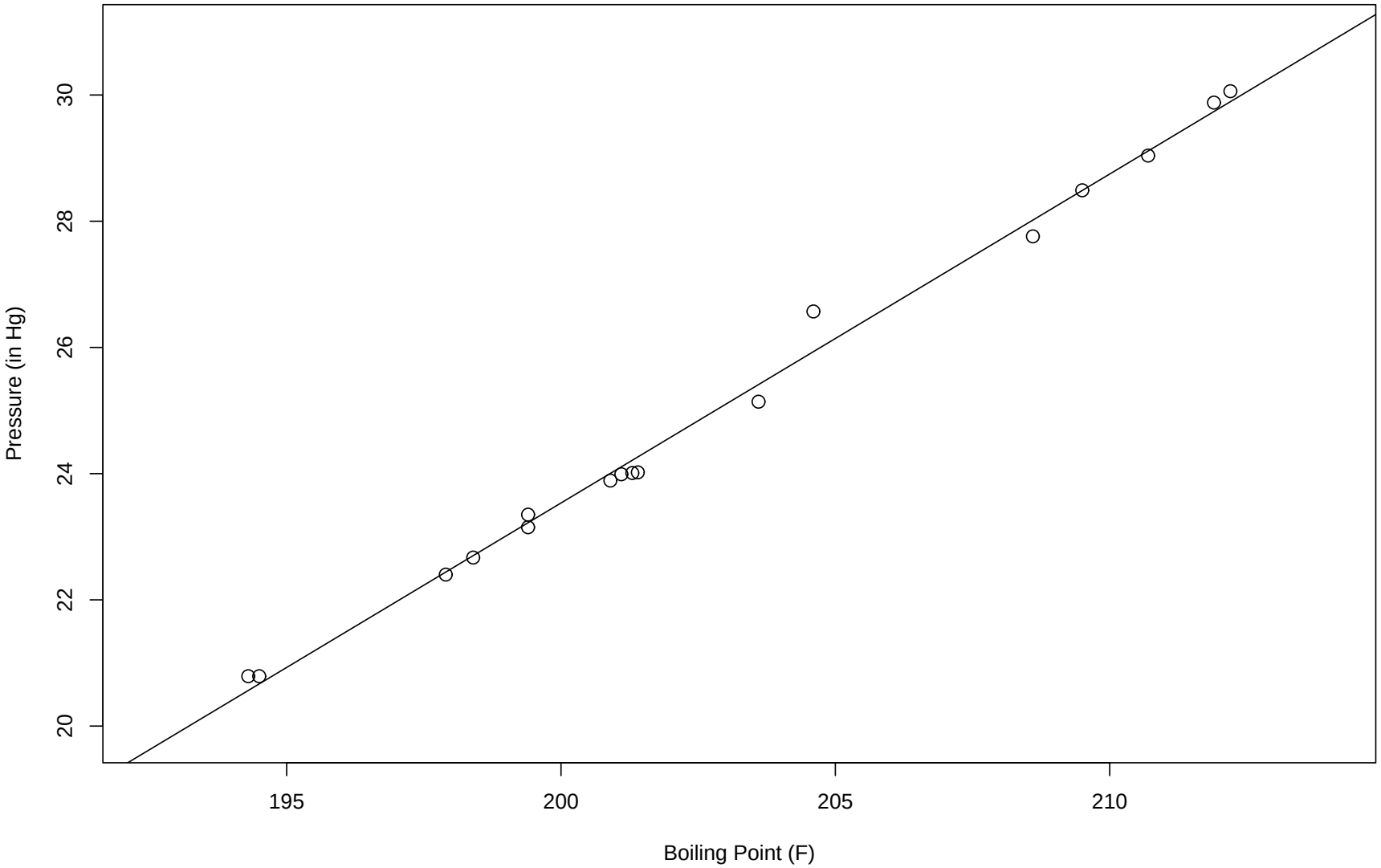
Example (Draper 1997): **Experimental data** from Weisberg (1985), gathered by the Scottish physicist **Forbes** in the mid-19th century ($n = 17$), relating the **boiling point** of water (x) to **barometric pressure** (y).

Boiling Point (°F)	194.5	194.3	197.9	198.4	199.4	199.4	$\cdots$	212.2
Pressure (in. Hg)	20.79	20.79	22.40	22.67	23.15	23.35	$\cdots$	30.06

Forbes theorized that log(pressure) should be **linear** in boiling point, but several **linear** and **nonlinear relationships** (**structural assumptions**) are **plausible** with these data:

$$
\begin{aligned}
S_1: \quad y_i &= \beta_{10} + \beta_{11} x_i + e_{1i}, & V(e_{1i}) &= \sigma_1^2, \\
S_2: \log(y_i) &= \beta_{20} + \beta_{21} x_i + e_{2i}, & V(e_{2i}) &= \sigma_2^2, \\
S_3: \quad y_i &= \beta_{30} + \beta_{31} \log(x_i) + e_{3i}, & V(e_{3i}) &= \sigma_3^2, \\
S_4: \log(y_i) &= \beta_{40} + \beta_{41} \log(x_i) + e_{4i}, & V(e_{4i}) &= \sigma_4^2.
\end{aligned}
\tag{9}
$$

BMA (continued)



BMA (continued)

If you were proceeding **empirically** in this problem, your **structural uncertainty** might be spanned by the set $\mathcal{S} = \{S_1, \ldots, S_4\}$, with **each element** in this set corresponding to a **different set of parameters**, e.g., $\theta_{S_1} = (\beta_{10}, \beta_{11}, \sigma_1), \ldots, \theta_{S_4} = (\beta_{40}, \beta_{41}, \sigma_4)$.

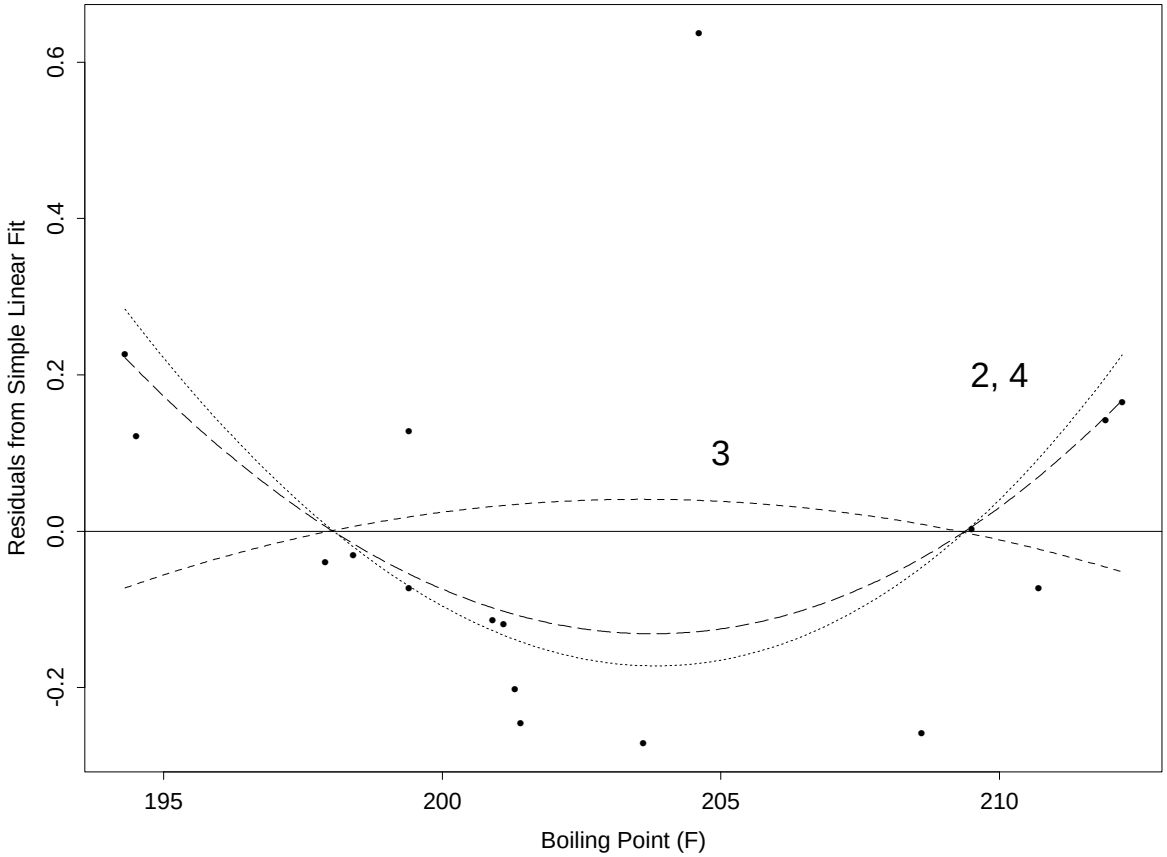
Suppose you're interested in **predicting** the y value (y^*, say) for a new $x = x^*$ at the **edge** of the observed data, as a way of helping to **discriminate** among these **four structural possibilities**.

Then, given the above **data set** D of (x, y) pairs, and **conditioning** on the set $\mathcal{S}$ of **structural possibilities**, the **predictive distribution** for y^* given x^* is

$$p(y^*|\mathcal{S}, x^*, D) = \sum_{j=1}^{m} p(y^*|S_j, x^*, D)\, p(S_j|D). \tag{10}$$

This is Bayesian model averaging: the **composite predictive distribution** $p(y^*|\mathcal{S}, x^*, D)$ is a **weighted average** of the **conditional predictive distributions** $p(y^*|S_j, x^*, D)$ given the **structural possibilities**, weighted by the **posterior probabilities** $p(S_j|D)$ of each of the **structural choices**.

BMA (continued)



Residual plot of the Forbes' Law data, with the **residuals** calculated from **structure 1** and with the **medians** of the **conditional predictive distributions** from **structures 2–4** superimposed: structures **2** and **4** (convex) **capture** the **nonlinearity** much better than structures **1** (horizontal) and **3** (concave).

BMA (continued)

$$p(y^*|\mathcal{S}, x^*, D) = \sum_{j=1}^{m} p(y^*|S_j, x^*, D)\, p(S_j|D)$$

Here the **conditional predictive distributions** $p(y^*|S_j, x^*, D)$ are just **scaled t distributions**, and

$$p(S_j|D) = c\, p(S_j)\, p(D|S_j), \tag{11}$$

where the $p(S_j)$ are the **prior structural probabilities** and

$$p(D|S_j) = \int_{\Theta_j} p(D|S_j, \theta_{S_j})\, p(\theta_{S_j}|S_j)\, d\theta_{S_j}. \tag{12}$$

Here $p(\theta_{S_j}|S_j)$ is the **prior distribution** for structure j's **parameter vector** and $p(D|S_j, \theta_{S_j})$ is the **likelihood function** for that structure.

I'll take $p(S_j)$ to be **constant** here to see what the **data** have to say; if in fact on **physical grounds** you believed in **Forbes' theory** in advance of the data, you would use a prior that **reflects** that **judgment**.

BMA (continued)

As for the $p(D|S_j)$, the ratio $\frac{p(D|S_1)}{p(D|S_2)}$ is just the **Bayes factor** for judging the **relative plausibility** of structure 1 against 2.

Various methods may be used to **approximate** the $p(D|S_j)$, including **MCMC** (e.g., Gilks et al., 1996) and **Laplace approximations** (e.g., Kass and Raftery, 1995; Draper, 1995).

Here the **simplest** of the latter type, based on the **Bayesian Information Criterion (BIC)** of Schwarz (1978), is adequate for my purposes:

$$\ln p(D|S_j) \doteq \ln p(D|\hat{\theta}_{S_j}, S_j) - \frac{1}{2} k_j \ln n, \tag{13}$$

where k_j and $\hat{\theta}_{S_j}$ are the **dimension** and **maximum likelihood estimate** of θ_{S_j}, respectively.

Since here **all four structures** have the **same number of parameters**, approximating $p(D|S_j)$ amounts in this case to **computing the maximum log likelihood value** for each structure and **normalizing**.

BMA (continued)

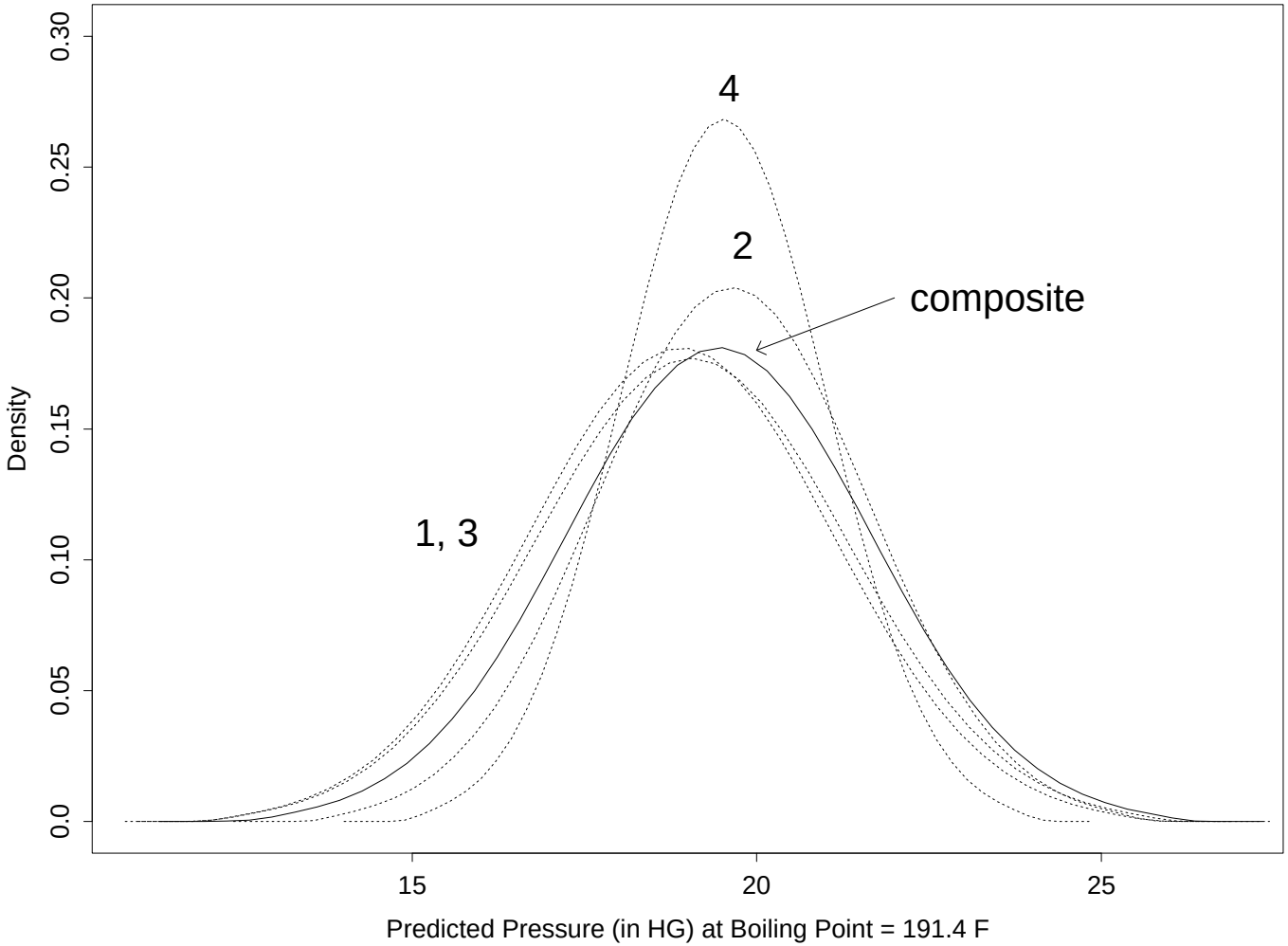
Maximum log likelihood values (column 2), **posterior structural probabilities** (column 3, based on BIC), and **means** and **standard deviations** of the **conditional** and **composite predictive distributions** with the Forbes data at $x^* = 191.4°\text{F}$ (columns 4 and 5):

Structure (j)	$\text{loglik}_{\max_j}$	$p(D\|S_j)$	Mean	SD
1 ($p - b$)	17.23	.130	19.05	0.285
2 ($\log(p) - b$)	17.89	.251	19.64	0.218
3 ($p - \log(b)$)	16.06	.040	18.90	0.309
4 ($\log(p) - \log(b)$)	18.73	.579	19.52	0.207
Composite	—	1.0	19.45	0.322

The **qualitative impression** conveyed by the **previous figure** is borne out by the **approximate posterior structural probabilities**, and there is **support** in this small data set for the **law proposed by Forbes**, although using p for pressure and b for boiling point, $\log(p)$-b and $\log(p)$-$\log(b)$ are **both** in the running.

BMA (continued)

Conditional predictive distributions for **structures 1–4** and the **composite predictive distribution**, for an x value **2 SDs below the mean**.



BMA (continued)

Structures **2** and **4** differ on average in their predictions by about **0.6inHg** from structures **1** and **3**, an amount that would be **discernible** in **new data** gathered to more strongly **discriminate** among the structures, but **a lot more data would be needed** to **disentangle** structures **2** and **4** (and in fact, **despite Forbes' conclusion**, the present small data set **supports** the **log(pressure)–log(boiling point)** model more strongly than **log(pressure)–boiling point**).

Note also that the **variance** of the **composite distribution, 0.104**, is about **twice as big** as the **probability-weighted average variance (0.0512)** of the four structures taken separately: this is **model uncertainty in action**.

BMA solves the **model uncertainty problem** by attempting to **honestly include** the **component of uncertainty** arising from **lack of perfect information** about the **underlying structure** of the **process** under study: the **uncertainty bands** again **widen** in pursuit of **better calibration**.