STOCHASTICALLY-ORDERED PARAMETERS IN BAYESIAN PREDICTION

by

WILLIAM S. JEWELL

UNIVERSITY OF CALIFORNIA · BERKELEY
STOCHASTICALLY-ORDERED PARAMETERS IN BAYESIAN PREDICTION

by

William S. Jewell
Department of Industrial Engineering
and Operations Research
University of California, Berkeley

OCTOBER 1979

ORC 79-12

This research has been partially supported by the Air Force Office of Scientific Research (AFSC), USAF, under Grants AFOSR-81-0122 and AFOSR-77-3179 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.
STOCHASTICALLY-ORDERED PARAMETERS IN BAYESIAN PREDICTION

Also supported by the Air Force Office of Scientific Research (AFOSR), USAF, under Grant AFOSR-77-3179.
ABSTRACT

In models of reliability growth in stages, it is usual to assume that system parameters improve monotonically from stage to stage, following some postulated law of growth. This paper explores a Bayesian model where such improvement only occurs on the average, e.g., a case when the parameters are assumed to be stochastically ordered. It is shown that the problem can be recast into a hierarchical form in which there are strictly-ordered hyperparameters which index the admissible family of ordered distributions for the parameters; the modelling problem is then to describe an appropriate law of motion over the hyperparameters.
0. INTRODUCTION

Models of reliability growth have received increasing attention in recent literature. One approach is to assume an underlying continuous improvement in system parameters; we call this the learning-curve approach (Jewell [2]). A second model, which may be more appropriate under certain circumstances, is growth by stages, in which the parameters increase (or decrease) monotonically at certain fixed points in time, when design improvements are made. In both of these models, an important problem is the prediction of the ultimate performance of the system after all reliability growth has occurred or all the design modifications have been effected; in practical procurement problems, it is desirable to make these predictions using as little test data as possible.

This paper explores an extension of the growth by stages model suggested by N. Singpurwalla [3] in which the parameters improve only stochastically, rather than absolutely.
I. DETERMINISTICALLY-ORDERED PARAMETERS

Consider the usual set-up, in which random observables \( x_1, x_2, \ldots, x_t \) in periods 1, 2, \ldots, \( t \) are governed by random parameters according to some likelihood density \( p(x_1, x_2, \ldots, x_t \mid \theta_1, \theta_2, \ldots, \theta_t) \). Given a prior parameter density \( p(\theta_1, \theta_2, \ldots, \theta_t) \), and observations \( x = (x_1, x_2, \ldots, x_t) \), a straightforward application of Bayes' law calculates the posterior parameter density \( p(\theta_1, \theta_2, \ldots, \theta_t \mid x) \).

In most applications, \( \hat{\theta}_t \) is assumed sufficient for \( x_t \), so that the likelihood takes the simpler form:

\[
(1.1) \quad p(x_1, x_2, \ldots, x_t \mid \theta_1, \theta_2, \ldots, \theta_t) = \prod_{i=1}^{t} p(x_i \mid \theta_i);
\]

we retain explicit dependence upon the time period primarily to permit stages of different duration, but not usually different failure laws in each interval.

In the usual growth-by-stages model, the improvement is expressed in terms of a deterministic ordering of the parameters, say:

\[
(1.2) \quad \theta_1 \geq \theta_2 \geq \ldots \geq \theta_t,
\]

and this is handled in an obvious way through the definition of the parameter prior and its support. A usual simplification is to fix some \( \theta_0 \), and then suppose that \( \hat{\theta}_{i+1} \) depends only upon \( \hat{\theta}_i \), and not upon other parameters or the index \( i \), so that:

\[
(1.3) \quad p(\theta_1, \theta_2, \ldots, \theta_t) = \prod_{i=1}^{t} p(\theta_i \mid \theta_{i-1}) \quad (\theta_0 \geq \theta_1 \geq \theta_2 \geq \theta_t).
\]
Typically a prototype parameter density \( g(u), 0 \leq u \leq 1 \), is selected for modelling convenience, giving a common form to the "shrinkage" of the parameters, viz.

\[
(1.4) \quad p(\theta_i | \theta_{i-1}) = \frac{1}{\theta_{i-1}} g\left(\frac{\theta_i}{\theta_{i-1}}\right) \quad (0 \leq \theta_i \leq \theta_{i-1})
\]

\((i = 1, 2, \ldots, t)\).

The actual Bayesian calculations are complicated, but present no conceptual difficulties; Smith [4] gives an example with \( p(x_t | \theta_t) \) binomial, and \( g(\cdot) \) uniform. Of course, (1.2), (1.3) and (1.4) represent rather strong beliefs about the underlying growth mechanism.
2. STOCHASTICALLY-ORDERED PARAMETERS

N. Singpurwalla [3] has suggested that (1.2) might reasonably be replaced by the weaker hypothesis that

\[ \tilde{\theta}_1 \preceq \tilde{\theta}_2 \preceq \ldots \preceq \tilde{\theta}_t. \]

This might mean, for example, that while the failure rate tended to decrease from stage to stage, such growth was by no means certain in a given realization. We now modify the basic model to handle this assumption.

(2.1) implies that the parameters are selected from some family \( F \) of priors which are stochastically ordered over the different stages. Let \( \alpha \) be a scalar hyperparameter which indexes the members of this family. Then we can associate \( \alpha_1, \alpha_2, \ldots, \alpha_t \) with stages 1, 2, ..., \( t \), and arrange that

\[ \alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_t \]

(or an increasing version) selects an appropriately ordered distribution for each parameter. Specifically, \( F \) is a family of conditional prior densities \( p(\theta \mid \alpha) \), with complementary distribution \( F^c(\theta \mid \alpha) \), such that, for every value of \( \theta \) and every \( i < j \) (\( i, j = 1, 2, \ldots \))

\[ F^c(\theta \mid \alpha_i) \geq F^c(\theta \mid \alpha_j); \]

this can be visualized as in Figure 1. \( \alpha_t \) can be an actual hyperparameter of the (conditional) prior, or it can be an abstract index that selects one of an admissible member of ordered priors. The key point is to arrange things so that (2.2) guarantees (2.1).
FIGURE 1

ADMISSIBLE FAMILY, $F$, OF STOCHASTICALLY ORDERED CONDITIONAL PRIORS ON $\theta$
With (2.1) guaranteed, what can we say about the selection of the hyperparameters? In a Bayesian model we must now express our prior beliefs and previous experience about what values of \(a_1, a_2, \ldots, a_t\) might occur during the different stages. In other words, we consider the hyperparameters as random variables with a given hyperprior density \(p(a_1, a_2, \ldots, a_t)\) over the family \(F\) satisfying (2.2).

Effectively, the stochastic ordering has introduced a hierarchical Bayesian formulation (see, e.g., Jewell, [1]), in which we again have deterministic ordering, this time in the hyperparameters. Note also that this approach avoids the "pseudo-Bayesian" formulation described by Singpurwalla [3].

For convenience in the sequel, we assume a Markovian hyperprior similar to (1.3), so that, given \(a_0\), and a family of (hyperprior) transition densities \(p(a_i | a_{i-1})\), we have:

\[
P(a_1, a_2, \ldots, a_t) = \prod_{i=1}^{t} p(a_i | a_{i-1}) \quad (a_0 > a_1 > a_2 > \ldots > a_t).
\]

\(a_0\) could, of course, have an initial density of its own. But, for simplicity, we omit further explicit mention of \(a_0\).

In summary, we have the following model:

(a) Prior beliefs or experience specify:

(i) \(a_0\) and the hyperprior transition densities \(p(a_i | a_{i-1})\) with deterministic ordering (2.2);

(ii) the family \(F\) of admissible conditional priors, so that, given \(a_i, \theta_i\) has complementary distribution \(p^c(\theta_i | a_i)\) and stochastic ordering (2.1);
(b) A known failure law governs the different observational likelihoods $p_i(x_i \mid \theta_i)$.

Note that $\tilde{a}_i$ is sufficient for $\tilde{\theta}_i$ which is, in turn sufficient for $\tilde{x}_i$, so that, given $a = (a_1, \ldots, a_t)$, the problem of predicting $\tilde{\theta} = (\theta_1, \theta_2, \ldots, \theta_t)$ and/or $x = (x_1, x_2, \ldots, x_t)$ decomposes into several separate problems. However, the usual situation is that we are given the observations $x$, and want to update the distributions of the unobserved $\tilde{\theta}$ and $\tilde{a}$, or to predict future observations $\tilde{x}_{t+1}, \tilde{x}_{t+2}, \ldots$. Because of the Markovian dependence of the $\tilde{a}_i$, both of these problems require extensive applications of Bayes' law.
3. MARGINAL DENSITIES

As a preliminary, we indicate how to calculate the various marginal densities. For the parameters, the joint marginal is:

\[(3.1) \quad p(\theta_1, \theta_2, \ldots, \theta_t) = \iint \ldots \int \prod_{i=1}^{t} p(\theta_i | \alpha_i) p(\alpha_i | \alpha_{i-1}) d\alpha_i .\]

Perhaps the easiest way to calculate is by defining:

\[(3.2) \quad p(\theta_1; \alpha_2) = \int p(\theta_1 | \alpha_1) p(\alpha_1 | \alpha_0) d\alpha_1 ,\]

and then proceed iteratively to find:

\[(3.3) \quad p(\theta_1, \theta_2, \ldots, \theta_i; \alpha_{i+1}) = \int p(\theta_i | \alpha_i) p(\alpha_{i+1} | \alpha_i) \cdot p(\theta_1, \theta_2, \ldots, \theta_{i-1}; \alpha_i) d\alpha_i \]

for \( i = 2, 3, \ldots \). At any point \( t \) where the marginal density is desired,

\[(3.4) \quad p(\theta_1, \theta_2, \ldots, \theta_t) = \int p(\theta_t | \alpha_t) \cdot p(\theta_1, \theta_2, \ldots, \theta_{t-1}; \alpha_t) d\alpha_t .\]

It is clear that the \( \alpha_i \)'s constitute \( x \) s. hast. \( \alpha \) which is not Markovian (as in (1.3)). One can show, however, that the nature of \( F \), expressed in (2.3), implies ordered marginals:

\[(3.5) \quad \Pr \{ \tilde{\theta}_i > \theta \} \geq \Pr \{ \tilde{\theta}_j > \theta \} , \quad (i < j) \quad (\forall \theta) \]

as expected.
For the observables \( \{x_t\} \), the intermediate nature of the \( \{\theta_t\} \) suggests we define "bypass" conditional densities from the likelihood and prior, viz.,

\[
(3.6) \quad p_i(x_i | \alpha_i) = \int p_i(x_i | \theta_i)p(\theta_i | \alpha_i)d\theta_i
\]

for \( i = 1, 2, \ldots \). The joint marginal of the observations is then similar to (3.1):

\[
(3.7) \quad p(x_1, x_2, \ldots, x_t) = \int \cdots \int \prod_{i=1}^{t} p_i(x_i | \alpha_i)p(\alpha_i | \alpha_{i-1})d\alpha_i,
\]

which can also be "bootstrapped" as in (3.2), (3.3). Alternately, if (3.1) or (3.4) were first calculated,

\[
(3.8) \quad p(x_1, x_2, \ldots, x_t) = \int \cdots \int p(\theta_1, \theta_2, \ldots, \theta_t) \prod_{i=1}^{t} p_i(x_i | \theta_i)d\theta_i.
\]
4. UPDATING HYPERPARAMETERS

We turn now to the calculation of the inverse probabilities. By using the bypass conditional densities of (3.6), we find that (given $\alpha_0$):

$$p(\alpha_1, \ldots, \alpha_t; x_1, x_2, \ldots, x_t) = \prod_{i=1}^{t} p_i(x_i | \alpha_i) \cdot p(\alpha_i | \alpha_{i-1})$$

so that the update formula for the hyperparameters (i.e., the "hyperposterior-to-the-data" density) is:

$$p(\alpha_1, \ldots, \alpha_t | x) = \frac{\prod_{i=1}^{t} p_i(x_i | \alpha_i) \cdot p(\alpha_i | \alpha_{i-1})}{p(x)}$$

$p(x)$ is calculated in (3.7), but in most practical calculations, it is treated simply as a normalization factor. Note that "later" observables provide information about "earlier" hyperparameters because of the Markovian dependence.
5. PREDICTING HYPERPARAMETERS

The prediction of $\tilde{a}_{t+1}$ from $x$ is most easily "bootstrapped" in a manner similar to (3.2), (3.3). We define:

\begin{equation}
 p(x_1; a_2) = \int p_1(x_1 | a_1)p(a_2 | a_1)p(a_1 | a_0)da_1,
\end{equation}

and then proceed iteratively for $i = 2, 3, \ldots$.

\begin{equation}
 p(x_1, x_2, \ldots, x_i; a_{i+1}) = \int p_1(x_1 | a_1)p(a_{i+1} | a_i) \cdot p(x_1, x_2, \ldots, x_{i-1}; a_i)da_i.
\end{equation}

At any stage $t$, the prediction of $\tilde{a}_t$ is a simple normalization, viz.

\begin{equation}
 p(a_{t+1} | x) = \frac{p(x_1, x_2, \ldots, x_t; a_{t+1})}{p(x)}.
\end{equation}

The iterative calculations (5.1), (5.2) provide also a simple formula for updating the most recent hyperparameter:

\begin{equation}
 p(a_t | x) = \frac{p_t(x_t | a_t) \cdot p(x_1, x_2, \ldots, x_{t-1}; a_t)}{p(x)}.
\end{equation}
6. UPDATING PARAMETERS

It does not seem easy to provide updating for the parameters alone, since \( p(\theta_1, \ldots, \theta_t \mid x) \) requires a complex integration over the Markovian transition probabilities. If all the marginals (3.2), (3.3) have been calculated to give (3.4), then, of course,

\[
(6.1) \quad p(\theta_1, \theta_2, \ldots, \theta_t \mid x) = \frac{\prod_{i=1}^{t} p_t(x_t \mid \theta_t) p(\theta_1, \theta_2, \ldots, \theta_t)}{p(x)} .
\]

For the latest parameter only, the "bootstrap" in (5.2) can be used with

\[
(6.2) \quad p(\theta_t \mid x) = \frac{\int p_t(x_t \mid \theta_t) p(\theta_t \mid \alpha_t) p(x_1, x_2, \ldots, x_{t-1} ; \alpha_t) d\alpha_t}{p(x)} .
\]

All data must be used to update \( \theta_t \). However, if we also know \( \alpha_t \), then only the latest observation is sufficient, as can be seen from:

\[
(6.3) \quad p(\theta_t, \alpha_t \mid x) = \frac{p_t(x_t \mid \theta_t) p(\theta_t \mid \alpha_t)}{p(x_t \mid \alpha_t)} = p(\theta_t \mid \alpha_t, x_t) .
\]
7. OTHER PREDICTIONS

In fact, the hyperparameter forecast density (5.3) is the key to all other predictions, since, given $x$, only $\hat{a}_{t+1}$ is sufficient for the next period, since:

$$p(x_{t+1}, \theta_{t+1}, a_{t+1} | x)$$

(7.1)

$$= p_{t+1}(x_{t+1} | \theta_{t+1}) p(\theta_{t+1} | a_{t+1}) p(a_{t+1} | x);$$

marginal versions follow by integration.

For stages in the future, this generalizes to

$$p(x_{t+u}, \theta_{t+u}, a_{t+u} | x)$$

(7.2)

$$= p_{t+u}(x_{t+u} | \theta_{t+u}) p(\theta_{t+u} | a_{t+u})$$

$$\cdot \int p^{(u-1)}(\alpha_{t+u} | a_{t+1}) p(a_{t+1} | x) d\alpha_{t+1}$$

for $u = 2, 3, \ldots$, where $p^{(n)}(\cdot | \cdot)$ is the usual $n$-step Markovian transition probability.
8. MODEL VARIATIONS

The model described in (2.5) is actually quite general. First of all, it includes the deterministically-ordered parameter model of Section 1 by making the conditional prior, \( p(\theta | \alpha) \), a degenerate density—i.e., \( \theta_t \equiv \alpha_t \) for all \( t \); the transition densities of (2.4) are then those of (1.3).

If, on the other hand, we assume a time-invariant hyperparameter \( \tilde{\alpha} = \tilde{\alpha}_0 = \tilde{\alpha}_1 = \ldots = \tilde{\alpha}_t \), with initial hyperprior density \( \pi(\alpha) \), then the model reduces to a hypothesis mixing formulation, in which data from successive trials is helping us to select one of a (usually finite) number of priors \( p(\theta_t | \alpha) \). The parameter \( \tilde{\theta}_t \) is usually also stationary, and the calculations reduce to updating posterior to data mixing probabilities \( \pi(\alpha | \tilde{x}) \).

The strict ordering (2.2) only applies if there is a scalar hyperparameter in the prior, (or if it is an abstract index over \( F \)). Figure 2 shows the situation if the hyperparameter is a two-vector, \( \alpha_t = (\alpha_{1,t} \alpha_{2,t}) \). From the form of the conditional prior \( p(\theta | \alpha) \), we can in principle find the allowed region, \( R(\alpha_{t-1}) \), for \( \alpha_t \) such that \( \tilde{\theta}_t \preceq \tilde{\theta}_{t-1} \), as shown. This region depends upon \( \alpha_{t-1} \) and includes it; in other words, the stochastic ordering of the parameters induces a partial ordering over the vector-valued hyperparameters. The transition probabilities \( p(\alpha_t | \alpha_{t-1}) \) then provide the appropriate measure over \( R(\alpha_{t-1}) \).

We also see, upon reflection, that our formulation can produce many other forms of parameter relationships than (1.2) or (2.1), through appropriate choices of \( R \). For instance, if \( p(\theta | \alpha) \) is a
FIGURE 2

MAPPING OF HYPERPARAMETERS FOR STOCHASTIC ORDERING
gamma density, an ordering of the shape parameter, keeping the mean
fixed, will give a family $F$ of priors with ordered coefficients
of variation, ranging from exponential to degenerate densities.
Other variations are left to the reader's imagination.
9. APPLICATION

In spite of its generality, this approach has serious practical drawbacks. The first is that it is difficult to construct analytic models in which the hierarchical computations can be carried out explicitly; even with a natural conjugate prior on \( \theta \), there remains the problem of specifying the hyperparameter density in a convenient form. ([4] illustrates some of these problems using the shrinkage (1.4).)

Secondly, digital computations require an additional dimension of the discretized space, so that it is difficult to approximate the usual continuous densities. Limited computational experience with small, rectangular priors and hyperpriors indicated that the growth under stochastic ordering behaves as expected, with an occasional reversal of growth, but otherwise was not illuminating.

Finally, of course, we should question the hypothesis of stochastic ordering itself. A decision-maker might believe that improvement occurs "on the average", for instance, but would prefer to model the anti-growth situations in a completely different way, with, say, (2.1) replaced by an ordering of the means only. We doubt seriously that any real data could discriminate against two such competing models.
REFERENCES


