13
Informed priors

When building an empirical model we typically attempt to include our understanding of the phenomenon as part of the model. This commonly describes both classical and Bayesian analyses (usually with locally uninformed priors). However, what analysis can we undertake if we have no data (new evidence) on which to apply our model. The above modeling strategy leaves us in a quandary. With no new data, we are not (necessarily) in a state of complete ignorance and this setting suggests the folly of ignoring our background knowledge in standard data analysis. If our model building strategy adequately reflects our state of knowledge plus the new data, we expect inferences from the standard approach described above to match Bayesian inference based on our informed priors plus the new data. If not, we have been logically inconsistent in at least one of the analyses. Hence, at a minimum, Bayesian analysis with informed priors serves as a consistency check on our analysis.

In this section, we briefly discuss maximum entropy priors conditional on our state of knowledge (see Jaynes [2003]). Our state of knowledge is represented by various averages of background knowledge (this includes means, variances, covariances, etc.). This is what we refer to as informed priors. The priors reflect our state of knowledge but no more; hence, maximum entropy conditional on what we know about the problem. Apparently, the standard in physical statistical mechanics for over a century.

13.1 Maximum entropy

What does it mean to be completely ignorant? If we know nothing, then we are unable to differentiate one event or state from another. If we are unable to differentiate events then our probability assignment consistent with this is surely that each event is equally likely. To suggest otherwise, presumes some deeper understanding. In order to deal with informed priors it is helpful to contrast with complete ignorance and its probability assignment. Maximum entropy priors are objective in the sense that two (or more) individuals with the same background knowledge assign the same plausibilities regarding a given set of propositions prior to considering new evidence.

Shannon's [1948] classical information theory provides a measure of our ignorance in the form of entropy. Entropy is defined as

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

where $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. This can be developed axiomatically from the following conditions.

Condition 13.1 *Some numerical measure $H_n(p_1, \ldots, p_n)$ of "state of knowledge" exists.*

Condition 13.2 *Continuity: $H_n(p_1, \ldots, p_n)$ is a continuous function of p_i.*[1]

Condition 13.3 *Monotonicity: $H_n(p_1, \ldots, p_n)$ is a monotone increasing function of n.*[2]

Condition 13.4 *Consistency: if there is more than one way to derive the value for $H_n(p_1, \ldots, p_n)$, they each produce the same answer.*

Condition 13.5 *Additivity:*[3]

$$\begin{aligned} H_n(p_1, \ldots p_n) &= H_r(p_1, \ldots p_r) + w_1 H_k\left(\frac{p_1}{w_1}, \ldots, \frac{p_k}{w_1}\right) \\ &\quad + w_2 H_m\left(\frac{p_{k+1}}{w_2}, \ldots, \frac{p_{k+m}}{w_2}\right) + \cdots \end{aligned}$$

Now, we sketch the arguments. Let

$$h(n) \equiv H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$

---

[1] Otherwise, an arbitrarily small change in the probability distribution could produce a large change in $H_n(p_1, \ldots, p_n)$.

[2] Monotonicity provides a sense of direction.

[3] For instance, $H_3(p_1, p_2, p_3) = H_2(p_1, q) + qH_2\left(\frac{p_2}{q}, \frac{p_3}{q}\right)$.

and

$$p_i = \frac{n_i}{\sum_{i=1}^{n} n_i}$$

for integers n_i. Then, combining the above with condition 13.5 implies

$$h\left(\sum_{i=1}^{n} n_i\right) = H\left(p_1, \ldots, p_n\right) + \sum_{i=1}^{n} p_i h\left(n_i\right)$$

Consider an example where $n = 3$, $n_1 = 3, n_2 = 4, n_3 = 2$,

$$\begin{aligned} h\left(9\right) &= H\left(\frac{3}{9}, \frac{4}{9}, \frac{2}{9}\right) + \frac{3}{9}h\left(3\right) + \frac{4}{9}h\left(4\right) + \frac{2}{9}h\left(2\right) \\ &= H\left(\frac{3}{9}, \frac{4}{9}, \frac{2}{9}\right) + \frac{3}{9}H\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) + \frac{4}{9}H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{2}{9}H\left(\frac{1}{2}, \frac{1}{2}\right) \\ &= H\left(\frac{1}{9}, \ldots, \frac{1}{9}\right) \end{aligned}$$

If we choose $n_i = m$ then the above collapses to yield

$$h\left(mn\right) = h\left(m\right) + h\left(n\right)$$

and apparently $h\left(n\right) = K \log n$, but since we're maximizing a monotone increasing function in p_i we can work with

$$h\left(n\right) = \log n$$

then

$$\begin{aligned} h\left(\sum_{i=1}^{n} n_i\right) &= H\left(p_1, \ldots, p_n\right) + \sum_{i=1}^{n} p_i h\left(n_i\right) \\ &= H\left(p_1, \ldots, p_n\right) + \sum_{i=1}^{n} p_i \log n_i \end{aligned}$$

Rewriting yields

$$H\left(p_1, \ldots, p_n\right) = h\left(\sum_{i=1}^{n} n_i\right) - \sum_{i=1}^{n} p_i \log n_i$$

Substituting $p_i \sum_i n_i$ for n_i yields

$$\begin{aligned} H(p_1, \ldots, p_n) &= h\left(\sum_{i=1}^{n} n_i\right) - \sum_{i=1}^{n} p_i \log\left(p_i \sum_i n_i\right) \\ &= h\left(\sum_{i=1}^{n} n_i\right) - \sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log\left(\sum_i n_i\right) \\ &= h\left(\sum_{i=1}^{n} n_i\right) - \sum_{i=1}^{n} p_i \log p_i - \log\left(\sum_i n_i\right) \end{aligned}$$

Since $h(n) = \log n, h\left(\sum_{i=1}^{n} n_i\right) = \log\left(\sum_i n_i\right)$, and we're left with Shannon's entropy measure

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$$

13.2 Complete ignorance

Suppose we know nothing, maximization of H subject to the constraints involves solving the following Lagrangian for $p_i, i = 1, \ldots, n$, and λ_0.[4]

$$-\sum_{i=1}^{n} p_i \log p_i - (\lambda_0 - 1)\left(\sum_{i=1}^{n} p_i - 1\right)$$

The first order conditions are

$$\begin{aligned} -\lambda_0 - \log(p_i) &= 0 \quad \text{for all } i \\ \sum_{i=1}^{n} p_i - 1 &= 0 \end{aligned}$$

Then, the solution is

$$\begin{aligned} p_i &= \exp[-\lambda_0] \quad \text{for all } i \\ \lambda_0 &= \log n \end{aligned}$$

In other words, as expected, $p_i = \frac{1}{n}$ for all i. This is the maximum entropy probability assignment.

[4] It's often convenient to write the Lagrange multiplier as $(\lambda_0 - 1)$.

13.3 A little background knowledge

Suppose we know a bit more. In particular, suppose we know the mean is F. Now, the Lagrangian is

$$-\sum_{i=1}^{n} p_i \log p_i - (\lambda_0 - 1)\left(\sum_{i=1}^{n} p_i - 1\right) - \lambda_1\left(\sum_{i=1}^{n} p_i f_i - F\right)$$

where f_i is the realized value for event i. The solution is

$$p_i = \exp\left[-\lambda_0 - f_i\lambda_1\right] \quad \text{for all } i$$

For example, $n = 3$, $f_1 = 1$, $f_2 = 2$, $f_3 = 3$, and $F = 2.5$, the maximum entropy probability assignment and multipliers are[5]

$$\begin{array}{ll} p_1 & 0.116 \\ p_2 & 0.268 \\ p_3 & 0.616 \\ \lambda_0 & 2.987 \\ \lambda_1 & -0.834 \end{array}$$

13.4 Generalization of maximum entropy principle

Suppose variable x can take on n different discrete values $(x_1, \ldots x_n)$ and our background knowledge implies there are m different functions of x

$$f_k(x), \quad 1 \le k \le m < n$$

and these have expectations given to us in our statement of the background knowledge

$$E\left[f_k(x)\right] = F_k = \sum_{i=1}^{n} p_i f_k(x_i), \quad 1 \le k \le m$$

The set of probabilities with maximum entropy that satisfy these m constraints can be identified by Lagrangian methods. As above, the solution is

$$p_i = \exp\left[-\lambda_0 - \sum_{j=1}^{m} \lambda_j f_j(x_i)\right] \quad \text{for all } i$$

and the sum of the probabilities is unity,

$$1 = \sum_{i=1}^{n} p_i = \exp\left[-\lambda_0\right] \sum_{i=1}^{n} \exp\left[-\sum_{j=1}^{m} \lambda_j f_j(x_i)\right]$$

[5]Of course, if $F = 2$ then $p_i = \frac{1}{3}$ and $\lambda_1 = 0$.

Now define a partition function

$$Z(\lambda_1, \ldots, \lambda_m) \equiv \sum_{i=1}^{n} \exp\left[-\sum_{j=1}^{m} \lambda_j f_j(x_i)\right]$$

and we have

$$1 = \exp[-\lambda_0] Z(\lambda_1, \ldots, \lambda_m)$$

which reduces to

$$\exp[\lambda_0] = Z(\lambda_1, \ldots, \lambda_m)$$

or

$$\lambda_0 = \log[Z(\lambda_1, \ldots, \lambda_m)]$$

Since the average value F_k equals the expected value of $f_k(x)$

$$F_k = \exp[-\lambda_0] \sum_{i=1}^{n} f_k(x_i) \exp\left[-\sum_{j=1}^{m} \lambda_j f_j(x_i)\right]$$

and

$$\begin{aligned} -\frac{\partial \log[Z(\lambda_1, \ldots, \lambda_m)]}{\partial \lambda_k} &= \frac{\sum_{i=1}^{n} f_k(x_i) \exp\left[-\sum_{j=1}^{m} \lambda_j f_j(x_i)\right]}{Z(\lambda_1, \ldots, \lambda_m)} \\ &= \exp[-\lambda_0] \sum_{i=1}^{n} f_k(x_i) \exp\left[-\sum_{j=1}^{m} \lambda_j f_j(x_i)\right] \end{aligned}$$

Therefore,[6]

$$F_k = -\frac{\partial \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_k}$$

[6]Return to the example with $n = 3$, $f_1(x_1) = 1$, $f_1(x_2) = 2$, $f_1(x_3) = 3$, and $F = 2.5$. The partition function is

$$Z(\lambda_1) = \exp[-f_1\lambda_1] + \exp[-f_2\lambda_1] + \exp[-f_3\lambda_1].$$

It is readily verified that $-\frac{\partial \log Z(\lambda_1)}{\partial \lambda_1} = F = 2.5$ on substituting the values of the multipliers.

The maximum value of entropy is

$$\begin{aligned} H_{\max} &= \max\left[-\sum_{i=1}^{n} p_i \log p_i\right] \\ &= \exp\left[-\lambda_0\right] \sum_{i=1}^{n} \exp\left[-\sum_{j=1}^{m} \lambda_j f_j\left(x_i\right)\right] \left(\lambda_0 + \sum_{j=1}^{m} \lambda_j f_j\left(x_i\right)\right) \\ &= \lambda_0 + \exp\left[-\lambda_0\right] \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j\left(x_i\right) \exp\left[-\sum_{j=1}^{m} \lambda_j f_j\left(x_i\right)\right] \\ &= \lambda_0 + \sum_{j=1}^{m} \lambda_j F_j \end{aligned}$$

To establish support for a global maximum, consider two possible probability distributions

$$\sum_{i=1}^{n} p_i = 1 \quad p_i \geq 0$$

and

$$\sum_{i=1}^{n} u_i = 1 \quad u_i \geq 0$$

Note

$$\log x \leq x - 1 \quad 0 \leq x < \infty$$

with equality if and only if $x = 1$. Accordingly,

$$\sum_{i=1}^{n} p_i \log \frac{u_i}{p_i} \leq \sum_{i=1}^{n} p_i \left(\frac{u_i}{p_i} - 1\right) = \sum_{i=1}^{n} \left(u_i - p_i\right) = 0$$

with equality if and only if $p_i = u_i$, $i = 1, \ldots, n$. Rewrite the left hand side in terms of entropy for p_i

$$\begin{aligned} \sum_{i=1}^{n} p_i \log \frac{u_i}{p_i} &= \sum_{i=1}^{n} p_i \log u_i - \sum_{i=1}^{n} p_i \log p_i \\ &= \sum_{i=1}^{n} p_i \log u_i + H\left(p_1, \ldots p_n\right) \end{aligned}$$

Substitution into the inequality and rearrangement yields

$$H\left(p_1, \ldots p_n\right) \leq 0 - \sum_{i=1}^{n} p_i \log u_i$$

or

$$H\left(p_1, \ldots p_n\right) \leq \sum_{i=1}^{n} p_i \log \frac{1}{u_i}$$

Let

$$u_i \equiv \frac{1}{Z(\lambda_1, \ldots, \lambda_m)} \exp\left[-\sum_{j=1}^{m} \lambda_j f_j(x_i)\right]$$

where the partition function $Z(\lambda_1, \ldots, \lambda_m)$ effectively serves as a normalizing factor. Now we can write the inequality

$$H(p_1, \ldots p_n) \leq \sum_{i=1}^{n} p_i \log \frac{1}{u_i}$$

as

$$H(p_1, \ldots p_n) \leq \sum_{i=1}^{n} p_i \left[\log Z(\lambda_1, \ldots, \lambda_m) + \sum_{j=1}^{m} \lambda_j f_j(x_i)\right]$$

or

$$H(p_1, \ldots p_n) \leq \log Z(\lambda_1, \ldots, \lambda_m) + \sum_{j=1}^{m} \lambda_j E[f_j(x_i)]$$

Since p_i can vary over all possible probability distributions and it attains its maximum only when

$$p_i = u_i \equiv \frac{1}{Z(\lambda_1, \ldots, \lambda_m)} \exp\left[-\sum_{j=1}^{m} \lambda_j f_j(x_i)\right]$$

we have a general derivation for the maximum entropy probability assignment subject to background knowledge $F_j, j = 1, \ldots, m$.

13.5 Discrete choice model as maximum entropy prior

From here we can provide a more rigorous argument for the frequent utilization of logistic regression when faced with discrete choice analysis. The logit model for discrete choice D conditional on (regime differences in) covariates X is

$$\begin{aligned} \Pr(D \mid X) &= \frac{1}{1 + \exp[-Y]} \\ &= \frac{1}{1 + \exp[-X\gamma]} \end{aligned}$$

but the basis for this specification is frequently left unanswered. Following Blower [2004], we develop this model specification from the maximum entropy principle.

Bayesian revision yields

$$\Pr(D \mid X) = \frac{\Pr(D, X)}{\Pr(X)}$$

and for treatment selection

$$\Pr(D=1 \mid X) = \frac{\Pr(D=1,X)}{\Pr(D=1,X)+\Pr(D=0,X)}$$

Rewrite this expression as

$$\Pr(D=1 \mid X) = \frac{1}{1+\frac{\Pr(D=0,X)}{\Pr(D=1,X)}}$$

The maximum entropy probability assignments, denoted $\hbar$, for the joint likelihoods, $\Pr(D=1,X)$ and $\Pr(D=0,X)$, are

$$\Pr(D=1,X,\hbar) = \frac{\exp\left[-\sum_{j=1}^{m}\lambda_j f_j(X_1)\right]}{Z(\lambda_1,\ldots,\lambda_m)}$$

and

$$\Pr(D=0,X,\hbar) = \frac{\exp\left[-\sum_{j=1}^{m}\lambda_j f_j(X_0)\right]}{Z(\lambda_1,\ldots,\lambda_m)}$$

The likelihood ratio is

$$\begin{aligned}\frac{\Pr(D=0,X,\hbar)}{\Pr(D=1,X,\hbar)} &= \frac{\exp\left[-\sum_{j=1}^{m}\lambda_j f_j(X_0)\right]}{\exp\left[-\sum_{j=1}^{m}\lambda_j f_j(X_1)\right]} \\ &= \exp[-Y]\end{aligned}$$

where

$$Y = -\sum_{j=1}^{m}\lambda_j\{f_j(X_1)-f_j(X_0)\}$$

Hence, we have the logistic regression specification as a maximum entropy probability assignment

$$\begin{aligned}\Pr(D=1 \mid X,\hbar) &= \frac{1}{1+\frac{\Pr(D=0,X,\hbar)}{\Pr(D=1,X,\hbar)}} \\ &= \frac{1}{1+\exp[-Y]}\end{aligned}$$

13.6 Continuous priors

Applying the principle of maximum entropy to continuous prior distributions is more subtle. We sketch Jaynes' [2003, ch. 12] limit arguments by taking the discrete expression of entropy

$$H^d = -\sum_{i=1}^{n} p_i \log p_i$$

to a continuous expression for entropy

$$H_\ell^c = -\int_a^b p\left(x \mid \Im\right) \log \frac{p\left(x \mid \Im\right)}{m\left(x\right)} dx$$

whose terms are defined below.

Let the number of discrete points $x_i, i = 1, \ldots, n$, become very numerous such that

$$\lim_{n\to\infty} \frac{1}{n} \left(\text{number of points in } a < x < b\right) = \int_a^b m\left(x\right) dx$$

and assume this is sufficiently well-behaved that adjacent differences tend to zero such that

$$\lim_{n\to\infty} n\left(x_{i+1} - x_i\right) = \frac{1}{m\left(x_i\right)}$$

The discrete probability distribution p_i goes into a continuous density, $p\left(x \mid \Im\right)$, with background knowledge, $\Im$, via the limiting form of

$$p_i = p\left(x_i \mid \Im\right)\left(x_{i+1} - x_i\right)$$

or utilizing the limit above

$$p_i \to p\left(x_i \mid \Im\right) \frac{1}{nm\left(x_i\right)}$$

Since

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{nm\left(x_i\right)} = \int_a^b dx$$

the limit of discrete entropy is

$$\begin{aligned}
H_\ell^d &\equiv \lim_{n\to\infty} H^d \\
&= -\lim_{n\to\infty} \sum_{i=1}^{n} p_i \log p_i \\
&= -\lim_{n\to\infty} \sum_{i=1}^{n} \frac{p\left(x_i \mid \Im\right)}{nm\left(x_i\right)} \log \frac{p\left(x_i \mid \Im\right)}{nm\left(x_i\right)} \\
&= -\int_a^b p\left(x \mid \Im\right) \log \frac{p\left(x \mid \Im\right)}{nm\left(x\right)} dx
\end{aligned}$$

The limit contains an infinite term, $\log n$. Normalize H_ℓ^d by subtracting this term and we have Jaynes' continuous measure of entropy

$$\begin{aligned} H_\ell^c &\equiv \lim_{n\to\infty}\left[H_\ell^d - \log n\right] \\ &= -\int_a^b p\left(x \mid \Im\right)\log\frac{p\left(x \mid \Im\right)}{m\left(x\right)}dx + \int_a^b p\left(x \mid \Im\right)\log\left(n\right)dx - \log n \\ &= -\int_a^b p\left(x \mid \Im\right)\log\frac{p\left(x \mid \Im\right)}{m\left(x\right)}dx \end{aligned}$$

Next, we revisit maximum entropy for continuous prior distributions.

13.6.1 Maximum entropy

The maximum entropy continuous prior is normalized

$$\int_a^b p\left(x \mid \Im\right)dx = 1$$

and is constrained by m mean values F_k for the various different functions $f_k\left(x\right)$ from our background knowledge

$$F_k = \int_a^b f_k\left(x\right)p\left(x \mid \Im\right)dx \quad k = 1, 2, \ldots, m$$

Treating $m\left(x\right)$ as known, the solution to the Lagrangian identifies the maximum entropy continuous prior

$$p\left(x \mid \Im\right) = \frac{m\left(x\right)\exp\left[\lambda_1 f_1\left(x\right) + \cdots + \lambda_m f_m\left(x\right)\right]}{Z\left(\lambda_1, \ldots, \lambda_m\right)}$$

where the partition function is

$$Z\left(\lambda_1, \ldots, \lambda_m\right) = \int_a^b m\left(x\right)\exp\left[\lambda_1 f_1\left(x\right) + \cdots + \lambda_m f_m\left(x\right)\right]dx$$

and the Lagrange multipliers are determined from

$$F_k = -\frac{\partial \log Z\left(\lambda_1, \ldots, \lambda_m\right)}{\partial \lambda_k} \quad k = 1, 2, \ldots, m$$

Then, with the maximum entropy prior in hand, our best estimate (by quadratic loss) of any other function of the parameters, say $q\left(x\right)$, is

$$E\left[q\left(x\right)\right] = \int_a^b q\left(x\right)p\left(x \mid \Im\right)dx$$

What is the role of the invariance measure, $m\left(x\right)$? First note what $m\left(x\right)$ buys us. Inclusion of $m\left(x\right)$ in the entropy measure of our state of knowledge means the entropy measure H_ℓ^c, partition function, Lagrange multipliers, and $E\left[q\left(x\right)\right]$

are invariant under a transformation of parameters, say $x \rightarrow y(x)$. What does this imply for ignorance priors? Suppose we only know $a < x < b$, then there are no multipliers and

$$\begin{aligned} p(x \mid \Im) &= \frac{m(x)\exp[0]}{\int_a^b m(x)\exp[0]\,dx} \\ &= \frac{m(x)}{\int_a^b m(x)\,dx} \end{aligned}$$

so that, except for normalizing constant $\frac{1}{\int_a^b m(x)dx}$, $m(x)$ is the prior distribution $p(x \mid \Im)$. Next, we briefly discuss use of transformation groups for resolving the invariance measure, $m(x)$, and fully specifying ignorance priors.

13.6.2 Transformation groups

We focus on ignorance priors since the maximum entropy principle dictates only our background knowledge is included in the prior; this means we must recognize our state of ignorance. Consider one of the most common problems in practice, a two parameter sampling distribution. We observe a sample $x_1, \ldots, x_n$ from a continuous sampling distribution $p(x \mid \nu, \sigma)\,dx = \phi(\nu, \sigma)\,dx$ where ν is a location parameter and σ is a scale parameter and we wish to estimate ν and σ. Suppose we have no knowledge of the location and scale parameters. What is the prior distribution $p(\nu, \sigma \mid \Im)\,d\nu d\sigma = f(\nu, \sigma)\,d\nu d\sigma$? What does it mean to have no knowledge of the location and scale parameters? Jaynes [2003, ch. 12] suggests the following characterization. If a change of location or scale alters our perception of the distribution of the parameters, we must not have been completely ignorant with regard to location and scale. Therefore, the distributions should be invariant to a transformation group.

Suppose we transform the variables as follows

$$\begin{aligned} \nu' &= \nu + b \\ \sigma' &= a\sigma \\ x' - \nu' &= a(x - \nu) \end{aligned}$$

$-\infty < b < \infty$ and $0 < a < \infty$. Invariance implies the sampling distribution for the transformed variables is the same as the sampling distribution for the original variables

$$p(x' \mid \nu', \sigma')\,dx' = \psi(x', \nu', \sigma')\,dx' = \phi(x, \nu, \sigma)\,dx$$

Similarly, the prior distribution for the transformed parameters, based on the Jacobian, is

$$g(\nu', \sigma') = a^{-1} f(\nu, \sigma)$$

These relations hold irrespective of the distributions $\phi(x, \nu, \sigma)$ and $f(\nu, \sigma)$.

If the sampling distribution is invariant under the above transformation group, then the two functions are the same

$$\psi(x,\nu,\sigma)=\phi(x,\nu,\sigma)$$

for all values a and b. Invariance to location and scale implies

$$\phi(x,\nu,\sigma)=\frac{1}{\sigma}h\left(\frac{x-\nu}{\sigma}\right)$$

for arbitrary function $h(\cdot)$.[7] Now, we return to priors.

Consider another problem with sample $x_1^{'},\ldots,x_n^{'}$ and we wish to estimate ν' and σ' but again have no initial knowledge of the location and scale. Let the prior distribution be $g(\nu',\sigma')$. Since we have two problems with the same background knowledge consistency requires we assign the same prior. Invariance to parameter transformation implies the functions are the same

$$f(\nu,\sigma)=g(\nu,\sigma)$$

Combining

$$g(\nu',\sigma')=a^{-1}f(\nu,\sigma)$$

with the transformation group gives

$$\begin{aligned} g(\nu+b,a\sigma) &= a^{-1}f(\nu,\sigma) \\ f(\nu,\sigma) &= ag(\nu+b,a\sigma) \end{aligned}$$

Now,

$$\begin{aligned} f(\nu,\sigma) &= g(\nu,\sigma) \\ f(\nu+b,a\sigma) &= g(\nu+b,a\sigma) \end{aligned}$$

combining this with the above yields

$$f(\nu,\sigma)=af(\nu+b,a\sigma)$$

Satisfying this condition implies the prior distribution is

$$f(\nu,\sigma)=\frac{\text{constant}}{\sigma}$$

— this is Jeffrey's prior.

To illustrate, suppose we only know $0<\nu<2$ and $1<\sigma<2$, then we can assign $m(\nu,\sigma)=\frac{1}{\sigma}$ and $f(\nu,\sigma)=\frac{1}{2\log 2}\frac{1}{\sigma}$. Now, consider the transformation $b=0.1$, and $a=\frac{1}{2}$, then $af(\nu+b,a\sigma)=\frac{1}{2}f\left(v+0.1,\frac{1}{2}\sigma\right)=\frac{1}{2\log 2}\frac{1}{2}\frac{1}{\frac{1}{2}\sigma}=\frac{1}{2\log 2}\frac{1}{\sigma}=f(\nu,\sigma)$ and $m(v',\sigma')=\frac{1}{2}\frac{1}{\sigma'}=\frac{1}{2}\frac{1}{\frac{1}{2}\sigma}=\frac{1}{\sigma}$. If we assign $m(v',\sigma')=\frac{1}{\sigma'}$, then $m(v,\sigma)=2\frac{1}{\sigma}=2\frac{1}{2\sigma'}=\frac{1}{\sigma'}$. The key is existence of $m(x)$.

[7] This discussion attempts to convey the intuitive implications of transformation groups for maximum entropy. See Jaynes [2003, p. 379] for a more complete discussion.

13.6.3 Uniform prior

Next, we temporarily suppress the invariance measure, $m(x)$, and derive a maximum entropy ignorance prior utilizing differential entropy

$$H = -\int_a^b f(x) \log f(x)\, dx$$

as a measure of continuous entropy. Suppose we're completely ignorant except that x has continuous support over the interval $\{a, b\}$. The maximum entropy prior distribution is surely uniform. Its derivation involves maximization of the limiting form of entropy such that $f(x) \geq 0$ and $\int_a^b f(x)\, dx = 1$. Following Cover and Thomas [1991, ch. 11], formulate the Lagrangian[8]

$$\mathcal{L} = -\int_a^b f(x) \log f(x)\, dx + \lambda_0 \left(\int_a^b f(x)\, dx - 1 \right)$$

Since the partial derivative of the functional $-\int_a^b f(x) \log f(x)\, dx$ with respect to $f(x)$ for each value x is

$$\begin{aligned} \frac{\partial}{\partial f(x_i)} \left[-\int_a^b f(x) \log f(x)\, dx \right] &= -\frac{\partial}{\partial f(x_i)} f(x_i) \log f(x_i) \\ &= -\log f(x_i) - 1 \end{aligned}$$

the gradient of the Lagrangian is

$$-\log f(x) - 1 + \lambda_0$$

Solving the first order conditions yields[9]

$$f(x) = \exp[-1 + \lambda_0]$$

Utilizing the constraint to solve for λ_0 we have

$$\begin{aligned} \int_a^b f(x)\, dx &= 1 \\ \int_a^b \exp[-1 + \lambda_0]\, dx &= 1 \\ \exp[-1 + \lambda_0](b - a) &= 1 \\ \lambda_0 &= 1 - \log(b - a) \end{aligned}$$

Now,

$$f(x) = \exp[-1 + \lambda_0]$$

---

[8] Alternatively, we could begin from the partition function.

[9] Since the second partial derivatives with respect to $f(x)$ are negative for all x, $-\frac{1}{f(x)}$, a maximum is assured.

becomes

$$\begin{aligned} f(x) &= \exp[-1+1-\log(b-a)] \\ f(x) &= \frac{1}{b-a} \end{aligned}$$

The maximum entropy prior with no background knowledge (other than continuity and support) is the uniform distribution. If we return to Jaynes' definition of continuous entropy then we can assign $m(x)=1$ (an invariance measure exists) and normalization produces $f(x)=\frac{m(x)}{\int_a^b m(x)dx}=\frac{1}{b-a}$, as discussed earlier. Hereafter, we work with differential entropy (for simplicity) and keep in mind the existence of $m(x)$.

13.6.4 Gaussian prior

Suppose our background knowledge is limited to a continuous variable with finite mean μ and finite variance σ^2. Following the development above, the Lagrangian is

$$\begin{aligned} \mathcal{L} &= -\int_{-\infty}^{\infty} f(x)\log f(x)\,dx + \lambda_0\left(\int_{-\infty}^{\infty} f(x)\,dx - 1\right) \\ &\quad +\lambda_1\left(\int_{-\infty}^{\infty} xf(x)\,dx - \mu\right) + \lambda_2\left(\int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx - \sigma^2\right) \end{aligned}$$

The first order conditions are

$$-1-\log f(x)+\lambda_0+\lambda_1 x+\lambda_2(x-\mu)^2=0$$

or

$$f(x)=\exp\left[-1+\lambda_0+\lambda_1 x+\lambda_2(x-\mu)^2\right]$$

Utilizing the constraints to solve for the multipliers involves

$$\begin{aligned} \int_{-\infty}^{\infty}\exp\left[-1+\lambda_0+\lambda_1 x+\lambda_2(x-\mu)^2\right]dx &= 1 \\ \int_{-\infty}^{\infty} x\exp\left[-1+\lambda_0+\lambda_1 x+\lambda_2(x-\mu)^2\right]dx &= \mu \\ \int_{-\infty}^{\infty}(x-\mu)^2\exp\left[-1+\lambda_0+\lambda_1 x+\lambda_2(x-\mu)^2\right]dx &= \sigma^2 \end{aligned}$$

A solution is[10]

$$\begin{aligned} \lambda_0 &= 1-\frac{1}{4}\log\left[4\pi^2\sigma^4\right] \\ \lambda_1 &= 0 \\ \lambda_2 &= -\frac{1}{2\sigma^2} \end{aligned}$$

[10]The result, $\lambda_1=0$, suggests how pivotal variance knowledge is to a Gaussian maximum entropy prior. In fact, for a given variance, the Gaussian distribution has maximum entropy.

Substitution of these values for the multipliers reveals

$$\begin{aligned} f(x) &= \exp\left[-1+\lambda_0+\lambda_1 x+\lambda_2 (x-\mu)^2\right] \\ f(x) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right] \end{aligned}$$

Hence, the maximum entropy prior given knowledge of the mean and variance is the Gaussian or normal distribution.

13.6.5 Multivariate Gaussian prior

If multiple variables or parameters are of interest and we have background knowledge of only their means μ and variances σ^2, then we know the maximum entropy prior for each is Gaussian (from above). Further, since we have no knowledge of their interactions, their joint prior is the product of the marginals.

Now, suppose we have background knowledge of the covariances as well. A straightforward line of attack is to utilize the Cholesky decomposition to write the variance-covariance matrix Σ as $\Gamma\Gamma^T$. We may now work with the transformed data $z = \Gamma^{-1}x$, derive the prior for z, and then by transformation of variables identify priors for x. Of course, since the prior for z is the product of marginal Gaussian priors, as before,

$$\begin{aligned} f(z_1,\ldots,z_k) &= f(z_1)\cdots f(z_k) \\ &= (2\pi)^{-\frac{k}{2}}\prod_{i=1}^{k}\exp\left[-\frac{1}{2}\left(z_i-\Gamma^{-1}\mu_i\right)^2\right] \end{aligned}$$

where $f(z_i) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(z_i-\Gamma^{-1}\mu_i\right)^2\right]$, the transformation back to the vector $x = \Gamma z$ produces the multivariate Gaussian distribution

$$\begin{aligned} f(x) &= (2\pi)^{-\frac{k}{2}} J\exp\left[-\frac{1}{2}\left(\Gamma^{-1}x-\Gamma^{-1}\mu\right)^T\left(\Gamma^{-1}x-\Gamma^{-1}\mu\right)\right] \\ &= (2\pi)^{-\frac{k}{2}} J\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \end{aligned}$$

where J is the Jacobian of the transformation. Since $J = \left|\Gamma^{-1}\right| = |\Gamma|^{-1}$ and $\Sigma = \left(LD^{\frac{1}{2}}\right)\left(D^{\frac{1}{2}}L^T\right) = \Gamma\Gamma^T$ is positive definite, $|\Gamma|^{-1} = |\Sigma|^{-\frac{1}{2}}$ where L is a lower triangular matrix and D is a diagonal matrix. Now, the density can be written in standard form

$$f(x) = (2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

Hence, the maximum entropy prior when background knowledge is comprised only of means, variances, and covariances for multiple variables or parameters is the multivariate Gaussian distribution.

13.6.6 Exponential prior

Suppose we know the variable of interest has continuous but non-negative support and finite mean β. The Lagrangian is

$$\begin{aligned}\mathcal{L} &= -\int_0^\infty f(x)\log f(x)\,dx + \lambda_0\left(\int_0^\infty f(x)\,dx - 1\right) \\ &\quad +\lambda_1\left(\int_0^\infty x f(x)\,dx - \beta\right)\end{aligned}$$

The first order conditions are

$$-1 - \log f(x) + \lambda_0 + \lambda_1 x = 0$$

Solving for $f(x)$ produces

$$f(x) = \exp\left[-1 + \lambda_0 + \lambda_1 x\right]$$

Using the constraints to solve for the multipliers involves

$$\begin{aligned}\int_0^\infty \exp\left[-1 + \lambda_0 + \lambda_1 x\right]dx &= 1 \\ \int_0^\infty x\exp\left[-1 + \lambda_0 + \lambda_1 x\right]dx &= \beta\end{aligned}$$

and produces

$$\begin{aligned}\lambda_0 &= 1 - \log\beta \\ \lambda_1 &= -\frac{1}{\beta}\end{aligned}$$

Substitution of these multipliers identifies the prior

$$\begin{aligned}f(x) &= \exp\left[-1 + \lambda_0 + \lambda_1 x\right] \\ f(x) &= \frac{1}{\beta}\exp\left[-\frac{x}{\beta}\right]\end{aligned}$$

Hence, the maximum entropy prior is an exponential distribution with mean β.

13.6.7 Truncated exponential prior

If support is shifted to, say, (a, ∞) for $a > 0$ and the mean equals β, the maximum entropy prior is a "truncated" exponential distribution. The first order conditions continue to be

$$-1 - \log f(x) + \lambda_0 + \lambda_1 x = 0$$

Solving for $f(x)$ again produces

$$f(x) = \exp\left[-1 + \lambda_0 + \lambda_1 x\right]$$

But using the constraints to solve for the multipliers involves

$$\begin{aligned}\int_a^\infty \exp\left[-1+\lambda_0+\lambda_1 x\right] dx &= 1 \\ \int_a^\infty x \exp\left[-1+\lambda_0+\lambda_1 x\right] dx &= \beta\end{aligned}$$

and produces

$$\begin{aligned}\lambda_0 &= 1 - \frac{a}{a+\beta} - \log\left[\beta - a\right] \\ \lambda_1 &= \frac{1}{a-\beta}\end{aligned}$$

Substitution of these multipliers identifies the prior

$$\begin{aligned}f(x) &= \exp\left[-1+\lambda_0+\lambda_1 x\right] \\ f(x) &= \frac{1}{\beta - a}\exp\left[-\frac{x-a}{\beta-a}\right]\end{aligned}$$

Hence, the maximum entropy prior is a "truncated" exponential distribution with mean β.

13.6.8 Truncated Gaussian prior

Suppose our background knowledge consists of the mean and variance over the limited support region, say (a, ∞), the maximum entropy prior is the truncated Gaussian distribution. This is consistent with the property the Gaussian distribution has maximum entropy of any distribution holding the variance constant.

As an example suppose we compare a mean zero Gaussian with the exponential distribution with variance one (hence, $a = 0$ and the mean of the exponential distribution is also one). If the variance of the truncated Gaussian equals one, then the underlying untruncated Gaussian has variance $\sigma^2 = 2.752$.[11] Entropy for the

[11] A general expression for the moments of a truncated Gaussian is

$$\begin{aligned}E\left[x \mid a \le x < b\right] &= \mu + \frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}\sigma \\ Var\left[x \mid a \le x < b\right] &= \sigma^2 \left[\begin{array}{c} 1 + \frac{\frac{a-\mu}{\sigma}\phi\left(\frac{a-\mu}{\sigma}\right) - \frac{b-\mu}{\sigma}\phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \\ - \left(\frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}\right)^2 \end{array}\right]\end{aligned}$$

where $\phi(\cdot)$ is the standard normal density function and $\Phi(\cdot)$ is the standard normal cumulative distribution function. For the setting under consideration, we set the variance of the truncated distribution

exponential distribution is

$$\begin{aligned} H &= -\int_0^\infty \exp[-x] \log(\exp[-x])\, dx \\ &= \int_0^\infty x \exp[-x]\, dx = 1 \end{aligned}$$

Entropy for the truncated Gaussian distribution is

$$\begin{aligned} H &= -\int_0^\infty \frac{2}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{x^2}{\sigma^2}\right] \log\left(\frac{2}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{x^2}{\sigma^2}\right]\right) dx \\ &= -\int_0^\infty \frac{2}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{x^2}{\sigma^2}\right] \left[\log\left(\frac{2}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2}\frac{x^2}{\sigma^2}\right] dx \\ &= 1.232 \end{aligned}$$

As claimed, a truncated Gaussian distribution with the same variance has greater entropy.

13.7 Variance bound and maximum entropy

A deep connection between maximum entropy distributions and the lower bound of the sampling variance (often called the Cramer-Rao lower bound) can now be demonstrated. Consider a sample of n observations

$$x \equiv \{x_1, x_2, \ldots, x_n\}$$

with sampling distribution dependent on θ, $p(x \mid \theta)$. Let

$$u(x, \theta) \equiv \frac{\partial \log p(x \mid \theta)}{\partial \theta}$$

and

$$(f, g) = \int f(x)\, g(x)\, dx$$

---

equal to one (equal to the variance of the exponential)

$$1 = \sigma^2 \left[1 - \left(\frac{\phi(0)}{1 - \Phi(0)}\right)^2\right]$$

and solve for σ^2. The mean of the truncated normal distribution is

$$E[x \mid 0 < x < \infty] = 0 + \sigma \frac{\phi(0)}{1 - \Phi(0)} = 1.324$$

By the Schwartz inequality we have

$$(f,g)^2 \leq (f,f)(g,g)$$

or, writing it out,

$$\left[\int f(x)\, g(x)\, dx\right]^2 = \int f(x)\, f(x)\, dx \int g(x)\, g(x)\, dx$$

where equality holds if and only if $f(x) = qg(x)$, $q = \frac{(f,g)}{(g,g)}$ not a function of x but possibly a function of θ.[12]

Now, choose

$$f(x) = u(x,\theta)\sqrt{p(x \mid \theta)}$$

and

$$g(x) = (\beta(x) - E[\beta])\sqrt{p(x \mid \theta)}$$

then

$$\begin{aligned}(f,g) &= \int u(x,\theta)(\beta(x) - E[\beta])\, p(x \mid \theta)\, dx \\ &= E[\beta u] - E[\beta]\, E[u]\end{aligned}$$

---

[12] Clearly, $\int [f(x) - qg(x)]^2\, dx \geq 0$. Now, find q to minimize the integral. The first order condition is

$$\begin{aligned}0 &= \int [f(x) - qg(x)]\, g(x)\, dx \\ 0 &= \int f(x)\, f(x)\, dx - q\int g(x)\, g(x)\, dx\end{aligned}$$

solving for q gives

$$q = \frac{(f,g)}{(g,g)}$$

and the inequality becomes an equality

$$\begin{aligned}\left[\int \frac{(f,g)}{(g,g)} g(x)\, g(x)\, dx\right]^2 &\leq \int \left(\frac{(f,g)}{(g,g)}\right)^2 g(x)\, g(x)\, dx \int g(x)\, g(x)\, dx \\ \left(\frac{(f,g)}{(g,g)}\right)^2 \left[\int g(x)\, g(x)\, dx\right]^2 &= \left(\frac{(f,g)}{(g,g)}\right)^2 \int g(x)\, g(x)\, dx \int g(x)\, g(x)\, dx\end{aligned}$$

since

$$\begin{aligned} E[u] &= \int u(x,\theta)\, p(x \mid \theta)\, dx \\ &= \int \frac{\partial \log p(x \mid \theta)}{\partial \theta} p(x \mid \theta)\, dx \\ &= \frac{\partial}{\partial \theta}\left[\int p(x \mid \theta)\, dx\right] \\ &= \frac{\partial}{\partial \theta}[1] \\ E[u] &= 0 \end{aligned}$$

we have

$$(f,g) = E[\beta u]$$

We also have

$$\begin{aligned} (f,f) &= \int [u(x,\theta)]^2 p(x \mid \theta)\, dx \\ &= E\left[u^2\right] \\ &= Var[u] \end{aligned}$$

the latter from $E[u] = 0$, and

$$\begin{aligned} (g,g) &= \int (\beta(x) - E[\beta])^2 p(x \mid \theta)\, dx \\ &= Var[\beta] \end{aligned}$$

So the Schwartz inequality simplifies to

$$E[\beta u]^2 \leq Var[\beta]\, Var[u]$$

or

$$E[\beta u] \leq \sqrt{Var[\beta]\, Var[u]}$$

But

$$\begin{aligned} E[\beta u] &= \int \beta(x) \frac{\partial \log p(x \mid \theta)}{\partial \theta} p(x \mid \theta)\, dx \\ &= \int \beta(x) \frac{\partial p(x \mid \theta)}{\partial \theta} dx \\ &= \frac{dE[\beta]}{d\theta} \\ &= 1 + b'(\theta) \end{aligned}$$

where $b(\theta) = (E[\beta] - \theta)$, bias in the parameter estimate, and $b'(\theta) = \frac{\partial b(\theta)}{\partial \theta} = \frac{\partial E[\beta]}{\partial \theta} - 1$. This means the inequality can be rewritten as

$$\begin{aligned} Var[\beta] &\geq \frac{E[\beta u]^2}{Var[u]} \\ &\geq \frac{[1 + b'(\theta)]^2}{\int \left[\frac{\partial \log p(x|\theta)}{\partial \theta}\right]^2 p(x \mid \theta)\, dx} \end{aligned}$$

A change of parameters $(\theta \to \tau)$ where $q(\theta) = -\frac{\partial \tau}{\partial \theta}$ and substitution into $f = qg$ yields

$$\begin{aligned} \frac{\partial \log p(x \mid \theta)}{\partial \theta} \sqrt{p(x \mid \theta)} &= -\frac{\partial \tau}{\partial \theta} (\beta(x) - E[\beta]) \sqrt{p(x \mid \theta)} \\ \frac{\partial \log p(x \mid \theta)}{\partial \theta} &= -\frac{\partial \tau}{\partial \theta} (\beta(x) - E[\beta]) \end{aligned}$$

Now, integrate over θ

$$\begin{aligned} \int \frac{\partial \log p(x \mid \theta)}{\partial \theta} d\theta &= \int -\tau'(\theta) (\beta(x) - E[\beta])\, d\theta \\ \log p(x \mid \theta) &= -\tau(\theta)\beta(x) + \int \frac{\partial \tau}{\partial \theta} E[\beta]\, d\theta \\ &= -\tau(\theta)\beta(x) + \int E[\beta]\, d\tau + \text{constant} \end{aligned}$$

Notice $\int E[\beta]\, d\tau$ is a function of θ, call it $-\log Z(\tau)$. Also, the constant is independent of θ but may depend on x, call it $\log m(x)$. Substitution gives

$$\begin{aligned} \log p(x \mid \theta) &= -\tau(\theta)\beta(x) - \log Z(\tau) + \log m(x) \\ p(x \mid \theta) &= \frac{m(x)}{Z(\tau)} e^{-\tau(\theta)\beta(x)} \end{aligned}$$

This is the maximum entropy distribution with a constraint[13] fixing $E[\beta(x)]$ and $Z(\tau)$ is a normalizing constant such that

$$Z(\tau) = \int m(x) e^{-\tau(\theta)\beta(x)} dx$$

The significance of this connection merits deeper consideration. If the sampling distribution is a maximum entropy distribution then maximal efficiency is achievable in the squared error loss sense, that is, the Cramer-Rao lower bound for the sampling variance is achievable.[14] Bayesian inference consistently processes all information by combining the maximum entropy prior distribution and maximum entropy likelihood function or sampling distribution. This affirms the power of probability as logic (Jaynes [2003]).

13.8 An illustration: Jaynes' widget problem

Jaynes' widget problem is a clever illustration of informed priors (Jaynes [1963], [2003], ch. 14). A manager of a process that produces red (*R*), yellow (*Y*), and green (*G*) widgets must choose between producing *R*, *Y*, or *G* widgets as only 200 of one type of widgets per day can be produced. If this is all that is known (nearly complete ignorance), the manager is indifferent between *R*, *Y*, or *G*. Suppose the manager acquires some background knowledge. For illustrative purposes, we explore stages of background knowledge.

Stage 1: The manager learns the current stock of widgets: 100 red, 150 yellow, and 50 green. With only this background knowledge including no knowledge of the consequences, the manager intuitively chooses to produce green widgets.

Stage 2: The manager learns the average daily orders have been 50 red, 100 yellow, and 10 green widgets. With this background knowledge, the manager may intuitively decide to produce yellow widgets.

---

[13] The constraint is $E[\beta(x)] = -\frac{\partial \log Z(\tau)}{\partial \tau}$ as

$$E[\beta(x)] = \int \beta(x) \frac{m(x)}{Z(\tau)} e^{-\tau(\theta)\beta(x)} dx$$

and

$$\begin{aligned} -\frac{\partial \log Z(\tau)}{\partial \tau} &= -\frac{\partial \int m(x) e^{-\tau(\theta)\beta(x)} dx}{\partial \tau} \\ &= -\frac{1}{Z(\tau)} \int m(x) e^{-\tau(\theta)\beta(x)} (-\beta(x)) dx \\ &= \int \beta(x) \frac{m(x)}{Z(\tau)} e^{-\tau(\theta)\beta(x)} dx \end{aligned}$$

[14] See Jaynes [2003], p. 520 for exceptions. Briefly, if the sampling distribution does not have the form of a maximum entropy distribution either the lower bound is not achievable or the sampling distribution has discontinuities.

Table 13.1: Jaynes' widget problem: summary of background knowledge by stage

Stage	*R*	*Y*	*G*	Decision
1. in stock	100	150	50	*G*
2. aver. daily orders	50	100	10	*Y*
3. aver. individual order size	75	10	20	*R*
4. specific order	0	0	40	?

Stage 3: The manager learns the average order size has been 75 red, 10 yellow, and 20 green widgets. With this background knowledge, the manager may intuitively switch to producing red widgets.

Stage 4: The manager learns an emergency order for 40 green widgets is imminent. Now, what does the manager decide to produce? It seems common sense is not enough to guide the decision. We'll pursue a formal analysis but first we summarize the problem in table 13.1.

Of course, this is a decision theoretic problem where formally the manager (a) enumerates the states of nature, (b) assigns prior probabilities associated with states conditional on background knowledge, (c) updates beliefs via Bayesian revision (as this framing of the problem involves no new evidence, this step is suppressed), (d) enumerates the possible decisions (produce *R*, *Y*, or *G*), and (e) selects the expected loss minimizing alternative based on a loss function which incorporates background knowledge of consequences.

13.8.1 Stage 1 solution

The states of nature are the number of red, yellow, and green widgets ordered today. Let $n_1 = 0, 1, 2, \ldots$ be the number of red widgets ordered. Similarly, let n_2 and n_3 be the number of yellow and green widgets ordered. If this triple (n_1, n_2, n_3) is known the problem is likely trivial. The maximum entropy prior given only stage 1 background knowledge is

$$\max_{p(n_1,n_2,n_3)} \left\{ -\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\sum_{n_3=0}^{\infty} p\left(n_1,n_2,n_3\right)\log p\left(n_1,n_2,n_3\right) \right\}$$
$$s.t. \quad \textstyle\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\sum_{n_3=0}^{\infty} p\left(n_1,n_2,n_3\right) = 1$$

or solve the Lagrangian

$$\begin{aligned}\mathcal{L} &= -\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\sum_{n_3=0}^{\infty} p\left(n_1,n_2,n_3\right)\log p\left(n_1,n_2,n_3\right) \\ &\quad -\left(\lambda_0 - 1\right)\left(\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\sum_{n_3=0}^{\infty} p\left(n_1,n_2,n_3\right) - 1\right)\end{aligned}$$

The solution is the improper (uniform) prior

$$p\left(n_1,n_2,n_3\right) = \exp\left[-\lambda_0\right] \quad \text{for all } \left(n_1,n_2,n_3\right)$$

where $\lambda_0 = \lim_{n\to\infty} \log n$.

As we have no background knowledge of consequences, the loss function is simply

$$R(x) = \begin{array}{ll} x & x > 0 \\ 0 & x \leq 0 \end{array}$$

and the loss associated with producing red widgets (decision D_1) is

$$L(D_1; n_1, n_2, n_3) = R(n_1 - S_1 - 200) + R(n_2 - S_2) + R(n_3 - S_3)$$

where S_i is the current stock of widget $i = R, Y$, or G. Similarly, the loss associated with producing yellow widgets (decision D_2) is

$$L(D_2; n_1, n_2, n_3) = R(n_1 - S_1) + R(n_2 - S_2 - 200) + R(n_3 - S_3)$$

or green widgets (decision D_3) is

$$L(D_3; n_1, n_2, n_3) = R(n_1 - S_1) + R(n_2 - S_2) + R(n_3 - S_3 - 200)$$

Then, the expected loss for decision D_1 is

$$\begin{aligned} E[L(D_1)] &= \sum_{n_i} p(n_1, n_2, n_3) L(D_1; n_1, n_2, n_3) \\ &= \sum_{n_1=0}^{\infty} p(n_1) R(n_1 - S_1 - 200) \\ &\quad + \sum_{n_2=0}^{\infty} p(n_2) R(n_2 - S_2) \\ &\quad + \sum_{n_3=0}^{\infty} p(n_3) R(n_3 - S_3) \end{aligned}$$

Expected loss associated with decision D_2 is

$$\begin{aligned} E[L(D_2)] &= \sum_{n_1=0}^{\infty} p(n_1) R(n_1 - S_1) \\ &\quad + \sum_{n_2=0}^{\infty} p(n_2) R(n_2 - S_2 - 200) \\ &\quad + \sum_{n_3=0}^{\infty} p(n_3) R(n_3 - S_3) \end{aligned}$$

and for decision D_3 is

$$\begin{aligned} E\left[L\left(D_3\right)\right] &= \sum_{n_1=0}^{\infty} p\left(n_1\right) R\left(n_1 - S_1\right) \\ &\quad + \sum_{n_2=0}^{\infty} p\left(n_2\right) R\left(n_2 - S_2\right) \\ &\quad + \sum_{n_3=0}^{\infty} p\left(n_3\right) R\left(n_3 - S_3 - 200\right) \end{aligned}$$

Recognize $p\left(n_i\right) = p$ for all n_i, let b any arbitrarily large upper limit such that $p = \frac{1}{b}$, and substitute in the current stock values

$$\begin{aligned} E\left[L\left(D_1\right)\right] &= \sum_{n_1=0}^{b} pR\left(n_1 - 300\right) + \sum_{n_2=0}^{b} pR\left(n_2 - 150\right) \\ &\quad + \sum_{n_3=0}^{b} pR\left(n_3 - 50\right) \\ &= \frac{\left(b-300\right)\left(b-299\right)}{2b} + \frac{\left(b-150\right)\left(b-149\right)}{2b} \\ &\quad + \frac{\left(b-50\right)\left(b-49\right)}{2b} \\ &= \frac{114500 - 997b + 3b^2}{2b} \end{aligned}$$

$$\begin{aligned} E\left[L\left(D_2\right)\right] &= \sum_{n_1=0}^{b} pR\left(n_1 - 100\right) + \sum_{n_2=0}^{b} pR\left(n_2 - 350\right) \\ &\quad + \sum_{n_3=0}^{b} pR\left(n_3 - 50\right) \\ &= \frac{\left(b-100\right)\left(b-99\right)}{2b} + \frac{\left(b-350\right)\left(b-349\right)}{2b} \\ &\quad + \frac{\left(b-50\right)\left(b-49\right)}{2b} \\ &= \frac{134500 - 997b + 3b^2}{2b} \end{aligned}$$

$$\begin{aligned} E\left[L\left(D_3\right)\right] &= \sum_{n_1=0}^{b} pR\left(n_1-100\right)+\sum_{n_2=0}^{b} pR\left(n_2-150\right) \\ &\quad +\sum_{n_3=0}^{b} pR\left(n_3-250\right) \\ &= \frac{\left(b-100\right)\left(b-99\right)}{2b}+\frac{\left(b-150\right)\left(b-149\right)}{2b} \\ &\quad +\frac{\left(b-250\right)\left(b-249\right)}{2b} \\ &= \frac{94500-997b+3b^2}{2b} \end{aligned}$$

Since the terms involving b are identical for all decisions, expected loss minimization involves comparison of the constants. Consistent with intuition, the expected loss minimizing decision is D_3.

13.8.2 Stage 2 solution

For stage 2 we know the average demand for widgets. Conditioning on these three averages adds three Lagrange multipliers to our probability assignment. Following the discussion above on maximum entropy probability assignment we have

$$p\left(n_1,n_2,n_3\right)=\frac{\exp\left[-\lambda_1 n_1-\lambda_2 n_2-\lambda_3 n_3\right]}{Z\left(\lambda_1,\lambda_2,\lambda_3\right)}$$

where the partition function is

$$Z\left(\lambda_1,\lambda_2,\lambda_3\right)=\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\sum_{n_3=0}^{\infty}\exp\left[-\lambda_1 n_1-\lambda_2 n_2-\lambda_3 n_3\right]$$

factoring and recognizing this as a product of three geometric series yields

$$Z\left(\lambda_1,\lambda_2,\lambda_3\right)=\prod_{i=1}^{3}\left(1-\exp\left[-\lambda_i\right]\right)^{-1}$$

Since the joint probability factors into

$$p\left(n_1,n_2,n_3\right)=p\left(n_1\right)p\left(n_2\right)p\left(n_3\right)$$

we have

$$p\left(n_i\right)=\left(1-\exp\left[-\lambda_i\right]\right)\exp\left[-\lambda_i n_i\right] \qquad \begin{array}{c} i=1,2,3 \\ n_i=0,1,2,\ldots \end{array}$$

$E\left[n_i\right]$ is our background knowledge and from the above analysis we know

$$\begin{aligned} E\left[n_i\right] &= -\frac{\partial \log Z\left(\lambda_1,\lambda_2,\lambda_3\right)}{\partial \lambda_i} \\ &= \frac{\exp\left[-\lambda_i\right]}{1-\exp\left[-\lambda_i\right]} \end{aligned}$$

Manipulation produces

$$\exp\left[-\lambda_i\right] = \frac{E\left[n_i\right]}{E\left[n_i\right]+1}$$

substitution finds

$$\begin{aligned} p\left(n_i\right) &= \left(1-\exp\left[-\lambda_i\right]\right)\exp\left[-\lambda_i n_i\right] \\ &= \frac{1}{E[n_i]+1}\left(\frac{E[n_i]}{E[n_i]+1}\right)^{n_i} \quad n_i = 0,1,2,\ldots \end{aligned}$$

Hence, we have three exponential distributions for the maximum entropy probability assignment

$$\begin{aligned} p_1\left(n_1\right) &= \frac{1}{51}\left(\frac{50}{51}\right)^{n_1} \\ p_2\left(n_2\right) &= \frac{1}{101}\left(\frac{100}{101}\right)^{n_2} \\ p_3\left(n_3\right) &= \frac{1}{11}\left(\frac{10}{11}\right)^{n_3} \end{aligned}$$

Now, combine these priors with the uninformed loss function, say for the first component of decision D_1

$$\begin{aligned} \sum_{n_1=0}^{\infty} p\left(n_1\right) R\left(n_1-300\right) &= \sum_{n_1=300}^{\infty} p\left(n_1\right)\left(n_1-300\right) \\ &= \sum_{n_1=300}^{\infty} p\left(n_1\right) n_1 - \sum_{n_1=300}^{\infty} p\left(n_1\right) 300 \end{aligned}$$

By manipulation of the geometric series

$$\begin{aligned} \sum_{n_1=300}^{\infty} p\left(n_1\right) n_1 &= \left(1-\exp\left[-\lambda_1\right]\right) \\ &\quad \times \frac{\exp\left[-300\lambda_1\right]\left(300\exp\left[\lambda_1\right]-299\right)\exp\left[-\lambda_1\right]}{\left(1-\exp\left[-\lambda_1\right]\right)^2} \\ &= \frac{\exp\left[-300\lambda_1\right]\left(300\exp\left[\lambda_1\right]-299\right)}{\exp\left[\lambda_1\right]-1} \end{aligned}$$

and

$$\begin{aligned} \sum_{n_1=300}^{\infty} p\left(n_1\right) 300 &= 300\left(1-\exp\left[-\lambda_1\right]\right)\frac{\exp\left[-300\lambda_1\right]}{1-\exp\left[-\lambda_1\right]} \\ &= 300\exp\left[-300\lambda_1\right] \end{aligned}$$

Combining and simplifying produces

$$\begin{aligned}\sum_{n_1=300}^{\infty} p(n_1)(n_1-300) &= \frac{\exp[-300\lambda_1](300\exp[\lambda_1]-299)}{\exp[\lambda_1]-1} \\ &\quad -\frac{\exp[-300\lambda_1](300\exp[\lambda_1]-300)}{\exp[\lambda_1]-1} \\ &= \frac{\exp[-300\lambda_1]}{\exp[\lambda_1]-1}\end{aligned}$$

substituting $\exp[-\lambda_1] = \frac{E[n_1]}{E[n_1]+1} = \frac{50}{51}$ yields

$$\sum_{n_1=300}^{\infty} p(n_1)(n_1-300) = \frac{\left(\frac{50}{51}\right)^{300}}{\frac{51}{50}-1} = 0.131$$

Similar analysis of other components and decisions produces the following summary results for the stage 2 decision problem.

$$\begin{aligned}E[L(D_1)] &= \sum_{n_1=0}^{\infty} p(n_1)R(n_1-300) + \sum_{n_2=0}^{\infty} p(n_2)R(n_2-150) \\ &\quad + \sum_{n_3=0}^{\infty} p(n_3)R(n_3-50) \\ &= 0.131 + 22.480 + 0.085 = 22.70\end{aligned}$$

$$\begin{aligned}E[L(D_2)] &= \sum_{n_1=0}^{\infty} p(n_1)R(n_1-100) + \sum_{n_2=0}^{\infty} p(n_2)R(n_2-350) \\ &\quad + \sum_{n_3=0}^{\infty} p(n_3)R(n_3-50) \\ &= 6.902 + 3.073 + 0.085 = 10.06\end{aligned}$$

$$\begin{aligned}E[L(D_3)] &= \sum_{n_1=0}^{\infty} p(n_1)R(n_1-100) + \sum_{n_2=0}^{\infty} p(n_2)R(n_2-150) \\ &\quad + \sum_{n_3=0}^{\infty} p(n_3)R(n_3-250) \\ &= 6.902 + 22.480 + 4\times 10^{-10} = 29.38\end{aligned}$$

Consistent with our intuition, the stage 2 expected loss minimizing decision is produce yellow widgets.

13.8.3 Stage 3 solution

With average order size knowledge, we are able to frame the problem by enumerating more detailed states of nature. That is, we can account for not only total orders but also individual orders. A state of nature can be described as we receive u_1 orders for one red widget, u_2 orders for two red widgets, etc., we also receive v_y orders for y yellow widgets and w_g orders for g green widgets. Hence, a state of nature is specified by

$$\theta = \{u_1, \ldots, v_1, \ldots, w_1, \ldots\}$$

to which we assign probability

$$p(u_1, \ldots, v_1, \ldots, w_1, \ldots)$$

Today's total demands for red, yellow and green widgets are

$$n_1 = \sum_{r=1}^{\infty} r u_r, \quad n_2 = \sum_{y=1}^{\infty} y v_y, \quad n_3 = \sum_{g=1}^{\infty} g w_g$$

whose expectations from stage 2 are $E[n_1] = 50$, $E[n_2] = 100$, and $E[n_3] = 10$. The total number of individual orders for red, yellow, and green widgets are

$$m_1 = \sum_{r=1}^{\infty} u_r, \quad m_2 = \sum_{y=1}^{\infty} v_y, \quad m_3 = \sum_{g=1}^{\infty} w_g$$

Since we know the average order size for red widgets is 75, for yellow widgets is 10, and for green widgets is 20, we also know the average daily total number of orders for red widgets is $E[m_1] = \frac{E[n_1]}{75} = \frac{50}{75}$, for yellow widgets is $E[m_2] = \frac{E[n_2]}{10} = \frac{100}{10}$, and for green widgets is $E[m_3] = \frac{E[n_3]}{20} = \frac{10}{20}$.

Six averages implies we have six Lagrange multipliers and the maximum entropy probability assignment is

$$p(\theta) = \frac{\exp[-\lambda_1 n_1 - \mu_1 m_1 - \lambda_2 n_2 - \mu_2 m_2 - \lambda_3 n_3 - \mu_3 m_3]}{Z(\lambda_1, \mu_1, \lambda_2, \mu_2, \lambda_3, \mu_3)}$$

Since both the numerator and denominator factor, we proceed as follows

$$\begin{aligned} p(\theta) &= p(u_1, \ldots, v_1, \ldots, w_1, \ldots) \\ &= p_1(u_1, \ldots) p_2(v_1, \ldots) p_3(w_1, \ldots) \end{aligned}$$

where, for instance,

$$\begin{aligned} Z_1(\lambda_1, \mu_1) &= \sum_{u_1=0}^{\infty} \sum_{u_2=0}^{\infty} \cdots \exp[-\lambda_1 (u_1 + 2u_2 + 3u_3 + \cdots)] \\ &\quad \times \exp[-\mu_1 (u_1 + u_2 + u_3 + \cdots)] \\ &= \prod_{r=1}^{\infty} \frac{1}{1 - \exp[-r\lambda_1 - \mu_1]} \end{aligned}$$

Since

$$E[n_i] = -\frac{\partial \log Z_i(\lambda_i, \mu_i)}{\partial \lambda_i}$$

and

$$E[m_i] = -\frac{\partial \log Z_i(\lambda_i, \mu_i)}{\partial \mu_i}$$

we can solve for, say, λ_1 and μ_1 via

$$\begin{aligned} E[n_i] &= \frac{\partial}{\partial \lambda_1} \sum_{r=1}^{\infty} \log\left(1 - \exp\left[-r\lambda_1 - \mu_1\right]\right) \\ &= \sum_{r=1}^{\infty} \frac{r}{\exp\left[r\lambda_1 + \mu_1\right] - 1} \end{aligned}$$

and

$$\begin{aligned} E[m_i] &= \frac{\partial}{\partial \mu_1} \sum_{r=1}^{\infty} \log\left(1 - \exp\left[-r\lambda_1 - \mu_1\right]\right) \\ &= \sum_{r=1}^{\infty} \frac{1}{\exp\left[r\lambda_1 + \mu_1\right] - 1} \end{aligned}$$

The expressions for $E[n_i]$ and $E[m_i]$ can be utilized to numerically solve for λ_i and μ_i to complete the maximum entropy probability assignment (see Tribus and Fitts [1968]), however, as noted by Jaynes [1963, 2003], these expressions converge very slowly. We follow Jaynes by rewriting the expressions in terms of quickly converging sums and then follow Tribus and Fitts by numerically solving for λ_i and μ_i.[15]

For example, use the geometric series

$$\begin{aligned} E[m_1] &= \sum_{r=1}^{\infty} \frac{1}{\exp\left[r\lambda_1 + \mu_1\right] - 1} \\ &= \sum_{r=1}^{\infty} \sum_{j=1}^{\infty} \exp\left[-j\left(r\lambda_1 + \mu_1\right)\right] \end{aligned}$$

Now, evaluate the geometric series over r

$$\sum_{r=1}^{\infty} \sum_{j=1}^{\infty} \exp\left[-j\left(r\lambda_1 + \mu_1\right)\right] = \sum_{j=1}^{\infty} \frac{\exp\left[-j\left(\lambda_1 + \mu_1\right)\right]}{1 - \exp\left[-j\lambda_1\right]}$$

[15] Jaynes [1963] employs approximations rather than computer-based numerical solutions.

Table 13.2: Jaynes' widget problem: stage 3 state of knowledge

Widget	S	$E[n_i]$	$E[m_i]$	λ_i	μ_i
Red	100	50	$\frac{50}{75}$	0.0134	4.716
Yellow	150	100	$\frac{100}{10}$	0.0851	0.514
Green	50	10	$\frac{10}{20}$	0.051	3.657

This expression is rapidly converging (the first term alone is a reasonable approximation). Analogous geometric series ideas apply to $E[n_i]$

$$\begin{aligned} E[n_1] &= \sum_{r=1}^{\infty} \frac{r}{\exp[r\lambda_1 + \mu_1] - 1} \\ &= \sum_{r=1}^{\infty}\sum_{j=1}^{\infty} r \exp[-j(r\lambda_1 + \mu_1)] \\ &= \sum_{j=1}^{\infty} \frac{\exp[-j(\lambda_1 + \mu_1)]}{(1 - \exp[-j\lambda_1])^2} \end{aligned}$$

Again, this series is rapidly converging. Now, numerically solve for λ_i and μ_i utilizing knowledge of $E[n_i]$ and $E[m_i]$. For instance, solving

$$\begin{aligned} E[m_1] &= \frac{50}{75} = \sum_{j=1}^{\infty} \frac{\exp[-j(\lambda_1 + \mu_1)]}{1 - \exp[-j\lambda_1]} \\ E[n_1] &= 50 = \sum_{j=1}^{\infty} \frac{\exp[-j(\lambda_1 + \mu_1)]}{(1 - \exp[-j\lambda_1])^2} \end{aligned}$$

yields $\lambda_1 = 0.0134$ and $\mu_1 = 4.716$. Other values are determined in analogous fashion and all results are described in table 13.2.[16]

Gaussian approximation

The expected loss depends on the distribution of daily demand, n_i. We compare a Gaussian approximation based on the central limit theorem with the exact distribution for n_i. First, we consider the Gaussian approximation. We can write the

[16] Results are qualitatively similar to those reported by Tribus and Fitts [1968].

expected value for the number of orders of, say, size r as

$$\begin{aligned} E[u_r] &= \sum_{u_r=0}^{\infty} p_1(u_r)\, u_r \\ &= \sum_{u_r=0}^{\infty} \frac{\exp[-(r\lambda_1+\mu_1)\, u_r]}{Z(\lambda_1,\mu_1)} u_r \\ &= \sum_{u_r=0}^{\infty} \frac{\exp[-(r\lambda_1+\mu_1)\, u_r]}{\frac{1}{1-\exp[-r\lambda_1-\mu_1]}} u_r \\ &= (1-\exp[-r\lambda_1-\mu_1]) \frac{\exp[-r\lambda_1-\mu_1]}{(1-\exp[-r\lambda_1-\mu_1])^2} \\ &= \frac{1}{\exp[r\lambda_1+\mu_1]-1} \end{aligned}$$

and the variance of u_r as

$$Var[u_r] = E\left[u_r^2\right] - E[u_r]^2$$

$$\begin{aligned} E\left[u_r^2\right] &= \sum_{u_r=0}^{\infty} \frac{\exp[-(r\lambda_1+\mu_1)\, u_r]}{\frac{1}{1-\exp[-r\lambda_1-\mu_1]}} u_r^2 \\ &= \sum_{u_r=0}^{\infty} (1-\exp[-r\lambda_1-\mu_1]) \\ &\quad \times \frac{\exp[-(r\lambda_1+\mu_1)] + \exp[-2(r\lambda_1+\mu_1)]}{(1-\exp[-r\lambda_1-\mu_1])^3} \\ &= \frac{\exp[r\lambda_1+\mu_1]+1}{(\exp[r\lambda_1+\mu_1]-1)^2} \end{aligned}$$

Therefore,

$$Var[u_r] = \frac{\exp[r\lambda_1+\mu_1]}{(\exp[r\lambda_1+\mu_1]-1)^2}$$

Since n_1 is the sum of independent random variables

$$n_1 = \sum_{r=1}^{\infty} r u_r$$

the probability distribution for n_1 has mean $E[n_1] = 50$ and variance

$$\begin{aligned} Var[n_1] &= \sum_{r=1}^{\infty} r^2 Var[u_r] \\ &= \sum_{r=1}^{\infty} \frac{r^2 \exp[r\lambda_1+\mu_1]}{(\exp[r\lambda_1+\mu_1]-1)^2} \end{aligned}$$

Table 13.3: Jaynes' widget problem: stage 3 state of knowledge along with standard deviation

Widget	S	$E[n_i]$	$E[m_i]$	λ_i	μ_i	σ_i
Red	100	50	$\frac{50}{75}$	0.0134	4.716	86.41
Yellow	150	100	$\frac{100}{10}$	0.0851	0.514	48.51
Green	50	10	$\frac{10}{20}$	0.051	3.657	19.811

We convert this into the rapidly converging sum[17]

$$\begin{aligned}\sum_{r=1}^{\infty}\frac{r^2\exp[r\lambda_1+\mu_1]}{(\exp[r\lambda_1+\mu_1]-1)^2} &= \sum_{r=1}^{\infty}\sum_{j=1}^{\infty} jr^2\exp[-j(r\lambda_1+\mu_1)]\\ &= \sum_{j=1}^{\infty} j\frac{\exp[-j(\lambda_1+\mu_1)]+\exp[-j(2\lambda_1+\mu_1)]}{(1-\exp[-j\lambda])^3}\end{aligned}$$

Next, we repeat stage 3 knowledge updated with the numerically-determined standard deviation of daily demand, σ_i, for the three widgets in table 13.3.[18,19]

The central limit theorem applies as there are many ways for large values of n_i to arise.[20] Then the expected loss of failing to meet today's demand given current stock, S_i, and today's production, $P_i = 0$ or 200, is

$$\begin{aligned}&\sum_{n_i=1}^{\infty} p(n_i)\,R(n_i-S_i-P_i)\\ \approx\ &\frac{1}{\sqrt{2\pi}\sigma_i}\int_{S_i+P_i}^{\infty}(n_i-S_i-P_i)\exp\left[-\frac{1}{2}\frac{(n_i-E[n_i])^2}{\sigma_i^2}\right]dn_i\end{aligned}$$

Numerical evaluation yields the following expected unfilled orders conditional on decision D_i.

$$\begin{aligned}E[L(D_1)] &= 0.05+3.81+0.16 = 4.02\\ E[L(D_2)] &= 15.09+0.0+0.16 = 15.25\\ E[L(D_3)] &= 15.09+3.81+0.0 = 18.9\end{aligned}$$

Clearly, producing red widgets is preferred given state 3 knowledge based on our central limit theorem (Gaussian) approximation. Next, we follow Tribus and Fitts [1968] and revisit the expected loss employing exact distributions for n_i.

---

[17] For both variance expressions, $Var[u_r]$ and $Var[n_1]$, we exploit the idea that the converging sum $\sum_{j=1}^{\infty} j^2\exp[-jx] = \frac{\exp[-x]+\exp[-2x]}{(1-\exp[-x])^3}$.

[18] Jaynes [1963] employs the quite good approximation $Var[n_i] \approx \frac{2}{\lambda_i}E[n_i]$.

[19] Results are qualitatively similar to those reported by Tribus and Fitts [1968].

[20] On the other hand, when demand is small, say, $n_i = 2$, there are only two ways for this to occur, $u_1 = 2$ or $u_2 = 1$.

Exact distributions

We derive the distribution for daily demand given stage 3 knowledge, $p(n_r \mid \Im_3)$, from the known distribution of daily orders $p(u_1, \ldots \mid \Im_3)$ by appealing to Bayes' rule

$$\begin{aligned} p(n_r \mid \Im_3) &= \sum_{u_1=0}^{\infty} \sum_{u_2=0}^{\infty} \cdots p(n_r u_1 u_2 \ldots \mid \Im_3) \\ &= \sum_{u_1=0}^{\infty} \sum_{u_2=0}^{\infty} \cdots p(n_r \mid u_1 u_2 \ldots \Im_3)\, p(u_1 u_2 \ldots \mid \Im_3) \end{aligned}$$

We can write

$$p(n_r \mid u_1 u_2 \ldots \Im_3) = \delta\left(n_r - \sum_{j=1}^{\infty} j u_j\right)$$

where $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ otherwise. Using independence of u_i, we have

$$p(n_r \mid \Im_3) = \sum_{u_1=0}^{\infty} \sum_{u_2=0}^{\infty} \cdots \delta\left(n_r - \sum_{j=1}^{\infty} j u_j\right) \prod_{i=1}^{\infty} p(u_i \mid \Im_3)$$

Definition 13.1 *Define the z transform as follows. For $f(n)$ a function of the discrete variable n, the z transform $F(z)$ is*

$$F(z) \equiv \sum_{n=0}^{\infty} f(n) z^n \quad 0 \leq z \leq 1$$

Let $P(z)$ be the z transform of $p(n_r \mid \Im_3)$

$$\begin{aligned} P(z) &= \sum_{n_r=0}^{\infty} \sum_{u_1=0}^{\infty} \sum_{u_2=0}^{\infty} \cdots z^{n_r} \delta\left(n_r - \sum_{j=1}^{\infty} j u_j\right) \prod_{i=1}^{\infty} p(u_i \mid \Im_3) \\ &= \sum_{u_1=0}^{\infty} \sum_{u_2=0}^{\infty} \cdots z^{\sum_{j=1}^{\infty} j u_j} \prod_{i=1}^{\infty} p(u_i \mid \Im_3) \\ &= \sum_{u_1=0}^{\infty} \sum_{u_2=0}^{\infty} \cdots \prod_{i=1}^{\infty} p(u_i \mid \Im_3) z^{i u_i} \\ &= \prod_{i=1}^{\infty} \sum_{u_i=0}^{\infty} z^{i u_i} p(u_i \mid \Im_3) \end{aligned}$$

Substituting $p(u_i \mid \Im_3) = (1 - \exp[-i\lambda_1 - \mu_1]) \exp[-u_i (i\lambda_1 + \mu_1)]$ yields

$$P(z) = \prod_{i=1}^{\infty} (1 - \exp[-i\lambda_1 - \mu_1]) \prod_{i=1}^{\infty} \sum_{u_i=0}^{\infty} \left(z^i \exp[-i\lambda_1 - \mu_1]\right)^{u_i}$$

Since $P(0) = \prod_{i=1}^{\infty} (1 - \exp[-i\lambda_1 - \mu_1])$, we can write

$$P(z) = P(0) \prod_{i=1}^{\infty} \sum_{u_i=0}^{\infty} \left(z^i \exp[-i\lambda_1 - \mu_1]\right)^{u_i}$$

The first few terms in the product of sums is

$$\begin{aligned} \frac{P(z)}{P(0)} &= \prod_{i=1}^{\infty} \sum_{u_i=0}^{\infty} \left(z^i \exp[-i\lambda_1 - \mu_1]\right)^{u_i} \\ &= 1 + \left(ze^{-\lambda_1}\right) e^{-\mu_1} + \left(ze^{-\lambda_1}\right)^2 \left[e^{-\mu_1} + e^{-2\mu_1}\right] \\ &\quad + \left(ze^{-\lambda_1}\right)^3 \left[e^{-\mu_1} + e^{-2\mu_1} + e^{-3\mu_1}\right] + \cdots \end{aligned}$$

Or, write

$$\frac{P(z)}{P(0)} = \sum_{n=0}^{\infty} C_n \left(ze^{-\lambda_1}\right)^n$$

where the coefficients C_n are defined by $C_0 = 1$ and

$$C_n = \sum_{j=1}^{n} C_{j,n} e^{-j\mu_1}, \quad \sum_{i=1}^{\infty} u_i = j, \quad \sum_{i=1}^{\infty} i u_i = n$$

and

$$C_{j,n} = C_{j-1,n-1} + C_{j,n-j}$$

with starting values $C_{1,1} = C_{1,2} = C_{1,3} = C_{1,4} = C_{2,2} = C_{2,3} = C_{3,3} = C_{3,4} = C_{4,4} = 1$ and $C_{2,4} = 2$.[21]

Let $p_0 \equiv p(n = 0 \mid \Im_3)$. Then, the inverse transform of $P(z)$ yields the distribution for daily demand

$$p(n \mid \Im_3) = p_0 C_n e^{-n\lambda_1}$$

We utilize this expression for $p(n \mid \Im_3)$, the coefficients $C_n = \sum_{j=1}^{n} C_{j,n} e^{-j\mu_1}$, the recursion formula $C_{j,n} = C_{j-1,n-1} + C_{j,n-j}$, and the earlier-derived Lagrange multipliers to numerically derive the distributions for daily demand for red, yellow, and green widgets. The distributions are plotted in figure 13.1.

As pointed out by Tribus and Fitts, daily demand for yellow widgets is nearly symmetric about the mean while daily demand for red and green widgets is "hit

[21] $C_{j,j} = 1$ for all j and $C_{j,n} = 0$ for all $n < j$. See the appendix of Tribus and Fitts [1968] for a proof of the recursion expression.

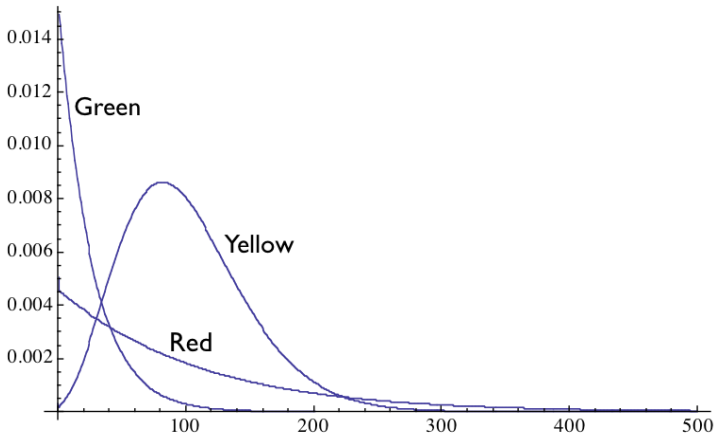

Figure 13.1: "Exact" distributions for daily widget demand

or miss." Probabilities of zero orders for the widgets are

$$\begin{aligned} p\left(n_1 = 0\right) &= 0.51 \\ p\left(n_2 = 0\right) &= 0.0003 \\ p\left(n_3 = 0\right) &= 0.61 \end{aligned}$$

Next, we recalculate the minimum expected loss decision based on the "exact" distributions. The expected loss of failing to meet today's demand given current stock, S_i, and today's production, $P_i = 0$ or 200, is

$$\sum_{n_i=1}^{\infty} p\left(n_i \mid \Im_3\right) R\left(n_i - S_i - P_i\right) = \sum_{S_i+P_i}^{\infty} \left(n_i - S_i - P_i\right) p\left(n_i \mid \Im_3\right)$$

Numerical evaluation yields the following expected unfilled orders conditional on decision D_i.

$$E\left[L\left(D_1\right)\right] = 2.35 + 5.07 + 1.31 = 8.73$$

$$E\left[L\left(D_2\right)\right] = 18.5 + 0.0 + 1.31 = 19.81$$

$$E\left[L\left(D_3\right)\right] = 18.5 + 5.07 + 0.0 = 23.58$$

While the Gaussian approximation for the distribution of daily widget demand and numerical evaluation of the "exact" distributions produce somewhat different expected losses, the both demonstrably support production of red widgets today.

13.8.4 Stage 4 solution

Stage 4 involves knowledge of an imminent order of 40 green widgets. This effectively changes the stage 3 analysis so that the current stock of green widgets is 10 rather than 50. Expected losses based on the Gaussian approximation are

$$E\left[L\left(D_1\right)\right] = 0.05 + 3.81 + 7.9 = 11.76$$

$$E\left[L\left(D_2\right)\right] = 15.09 + 0.0 + 7.9 = 22.99$$

$$E\left[L\left(D_3\right)\right] = 15.09 + 3.81 + 0.0 = 18.9$$

On the other hand, expected losses based on the "exact" distributions are

$$E\left[L\left(D_1\right)\right] = 2.35 + 5.07 + 6.70 = 14.12$$

$$E\left[L\left(D_2\right)\right] = 18.5 + 0.0 + 6.70 = 25.20$$

$$E\left[L\left(D_3\right)\right] = 18.5 + 5.07 + 0.0 = 23.58$$

While stage 4 knowledge shifts production in favor of green relative to yellow widgets, both distributions for daily widget demand continue to support producing red widgets today. Next, we explore another probability assignment puzzle.

13.9 Football game puzzle

Jaynes [2003] stresses consistent reasoning as the hallmark of the maximum entropy principle. Sometimes, surprisingly simple settings can pose a challenge. Consider the following puzzle posed by Walley [1991, pp. 270-271]. A football match-up between two football rivals produces wins (W), losses (L), or draws (D) for the home team. If this is all we know then the maximum entropy prior for the home team's outcome is uniform $\Pr\left(W, L, D\right) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. Suppose we know the home team wins half the time. Then, the maximum entropy prior is $\Pr\left(W, L, D\right) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$. Suppose we learn the game doesn't end in a draw. The posterior distribution is $\Pr\left(W, L, D\right) = \left(\frac{2}{3}, \frac{1}{3}, 0\right)$.[22]

Now, we ask what is the maximum entropy prior if the home team wins half the time and the game is not a draw. The maximum entropy prior is $\Pr\left(W, L, D\right) = \left(\frac{1}{2}, \frac{1}{2}, 0\right)$. What is happening? This appears to be inconsistent reasoning. Is there something amiss with the maximum entropy principle?

We suggest two different propositions are being evaluated. The former involves a game structure that permits draws but we gain new evidence that a particular game did not end in a draw. On the other hand, the latter game structure precludes draws. Consequently, the information regarding home team performance has a very different implication (three states of nature, W vs. L or D, compared with

[22] We return to this puzzle later when we discuss Jaynes' A_p distribution.

two states of nature, W vs. L). This is an example of what Jaynes [2003, pp. 470-3] calls "the marginalization paradox," where nuisance parameters are integrated out of the likelihood in deriving the posterior. If we take care to recognize these scenarios involve different priors and likelihoods, there is no contradiction. In Jaynes' notation where we let $\varsigma = W$, $y =$ not D, and $z =$ null, the former involves posterior $p(\varsigma \mid y, z, \Im_1)$ with prior $\Im_1$ permitting W, L, or D, while the latter involves posterior $p(\varsigma \mid z, \Im_2)$ with prior $\Im_2$ permitting only W or L. Evaluation of propositions involves joint consideration of priors and likelihoods, if either changes there is no surprise when our conclusions are altered.

The example reminds us of the care required in formulating the proposition being evaluated. The next example revisits an accounting issue where informed priors are instrumental to identification and inference.

13.10 Financial statement example

13.10.1 Under-identification and Bayes

If we have more parameters to be estimated than data, we often say the problem is under-identified. However, this is a common problem in accounting. To wit, we often ask what activities did the organization engage in based on our reading of their financial statements. We know there is a simple linear relation between the recognized accounts and transactions

$$Ay = x$$

where A is an $m \times n$ matrix of ± 1 and 0 representing simple journal entries in its columns and adjustments to individual accounts in its rows, y is the transaction amount vector, and x is the change in the account balance vector over the period of interest (Arya, et al [2000]). Since there are only $m - 1$ linearly independent rows (due to the balancing property of accounting) and m (the number of accounts) is almost surely less than n (the number of transactions we seek to estimate) we're unable to invert from x to recover y. Do we give up? If so, we might be forced to conclude financial statements fail even this simplest of tests.

Rather, we might take a page from physicists (Jaynes [2003]) and allow our prior knowledge to assist estimation of y. Of course, this is what decision theory also recommends. If our prior or background knowledge provides a sense of the first two moments for y, then the Gaussian or normal distribution is our maximum entropy prior. Maximum entropy implies that we fully utilize our background knowledge but don't use background knowledge we don't have (Jaynes [2003], ch. 11). That is, maximum entropy priors combined with Bayesian revision make efficient usage of both background knowledge and information from the data (in this case, the financial statements). As in previously discussed accounting examples, background knowledge reflects potential equilibria based on strategic interaction of various, relevant economic agents and accounting recognition choices for summarizing these interactions.

Suppose our background knowledge $\Im$ is completely summarized by

$$E\left[y \mid \Im\right] = \mu$$

and

$$Var\left[y \mid \Im\right] = \Sigma$$

then our maximum entropy prior distribution is

$$p\left(y \mid \Im\right) \sim N\left(\mu, \Sigma\right)$$

and the posterior distribution for transactions, y, conditional on the financial statements, x, is

$$\begin{array}{rl} & p\left(y \mid x, \Im\right) \\ \sim & N\left(\mu + \Sigma A_0^T \left(A_0 \Sigma A_0^T\right)^{-1} A_0 \left(y^p - \mu\right), \Sigma - \Sigma A_0^T \left(A_0 \Sigma A_0^T\right)^{-1} A_0 \Sigma\right) \end{array}$$

where $N\left(\cdot\right)$ refers to the Gaussian or normal distribution with mean vector denoted by the first term, and variance-covariance matrix denoted by the second term, A_0 is A after dropping one row and y^p is any consistent solution to $Ay = x$ (for example, form any spanning tree from a directed graph of $Ay = x$ and solve for y^p). For the special case where $\Sigma = \sigma^2 I$ (perhaps unlikely but nonetheless illuminating), this simplifies to

$$p\left(y \mid x, \Im\right) \sim N\left(P_{R(A)} y^p + \left(I - P_{R(A)}\right)\mu, \sigma^2\left(I - P_{R(A)}\right)\right)$$

where $P_{R(A)} = A_0^T \left(A_0 A_0^T\right)^{-1} A_0$ (projection into the rowspace of A), and then $I - P_{R(A)}$ is the projection into the nullspace of A.[23]

[23] In the general case, we could work with the subspaces (and projections) of $A_0\Gamma$ where $\Sigma = \Gamma\Gamma^T$ (the Cholesky decomposition of Σ) and the transformed data $z \equiv \Gamma^{-1} y \sim N\left(\Gamma^{-1}\mu, I\right)$ (Arya, Fellingham, and Schroeder [2000]). Then, the posterior distribution of z conditional on the financial statements x is

$$p\left(z \mid x, \Im\right) \sim N\left(P_{R(A_0\Gamma)} z^p + \left(I - P_{R(A_0\Gamma)}\right)\mu_z, I - P_{R(A_0\Gamma)}\right)$$

where $z^p = \Gamma^{-1} y^p$ and $\mu_z = \Gamma^{-1}\mu$. From this we can recover the above posterior distribution of y conditional on x via the inverse transformation $y = \Gamma z$.

13.10.2 Numerical example

Suppose we observe the following financial statements.

Balance sheets	Ending balance	Beginning balance
Cash	110	80
Receivables	80	70
Inventory	30	40
Property & equipment	110	100
Total assets	330	290
Payables	100	70
Owner's equity	230	220
Total equities	330	290

Income statement	for period
Sales	70
Cost of sales	30
SG&A	30
Net income	10

Let x be the change in account balance vector where credit changes are negative. The sum of x is zero; a basis for the left nullspace of A is a vector of ones.

change in account	amount
Δ cash	30
Δ receivables	10
Δ inventory	(10)
Δ property & equipment	10
Δ payables	(30)
sales	(70)
cost of sales	30
sg&a expenses	30

We envision the following transactions associated with the financial statements and are interested in recovering their magnitudes y.

transaction	amount
collection of receivables	y_1
investment in property & equipment	y_2
payment of payables	y_3
bad debts expense	y_4
sales	y_5
depreciation - period expense	y_6
cost of sales	y_7
accrued expenses	y_8
inventory purchases	y_9
depreciation - product cost	y_{10}

A crisp summary of these details is provided by a directed graph as depicted in figure 13.2.

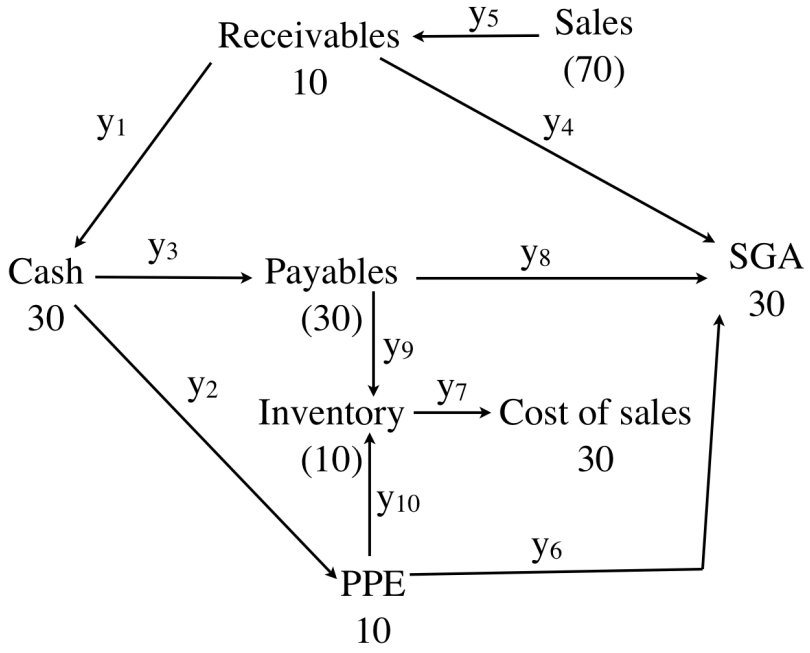

Figure 13.2: Directed graph of financial statements

The A matrix associated with the financial statements and directed graph where credits are denoted by -1 is

$$A = \begin{bmatrix} 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$

and a basis for the nullspace is immediately identified by any set of linearly independent loops in the graph, for example,

$$N = \begin{bmatrix} 1 & 0 & 1 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 1 & -1 \end{bmatrix}$$

A consistent solution y^p is readily identified by forming a spanning tree and solving for the remaining transaction amounts. For instance, let $y_3 = y_6 = y_9 = 0$, the spanning tree is depicted in figure 13.3.

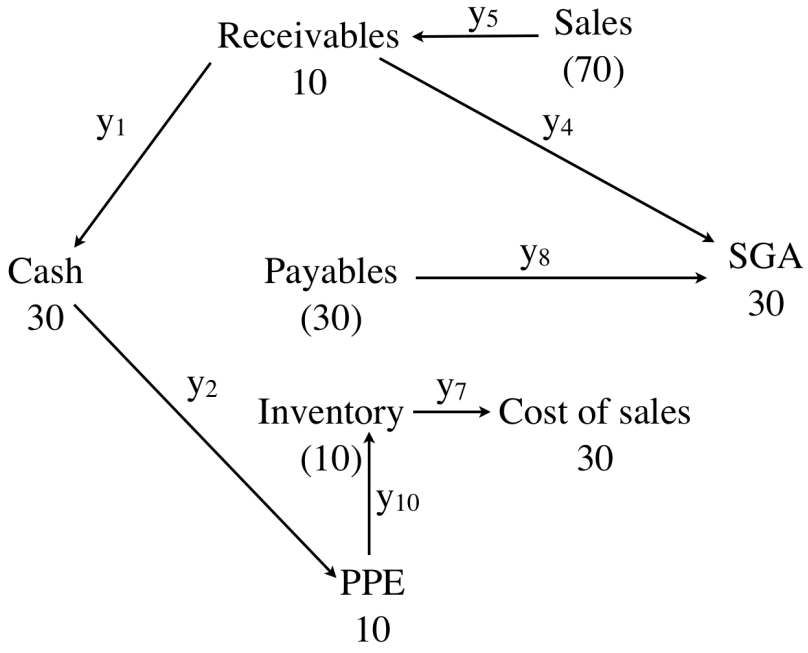

Figure 13.3: Spanning tree

Then, $(y^p)^T = \begin{bmatrix} 60 & 30 & 0 & 0 & 70 & 0 & 30 & 30 & 0 & 20 \end{bmatrix}$.

Now, suppose background knowledge $\Im$ regarding transactions is described by the first two moments

$$E\left[y^T \mid \Im\right] = \mu^T = \begin{bmatrix} 60 & 20 & 25 & 2 & 80 & 5 & 40 & 10 & 20 & 15 \end{bmatrix}$$

and

$$Var\left[y \mid \Im\right] = \Sigma = \begin{bmatrix} 10 & 0 & 0 & 0 & 5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0.2 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.1 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0.1 & 10 & 0 & 3.5 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.5 & 0 & 5 & 0 & 0.2 & 0 \\ 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 1 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

maximum entropy priors for transactions are normally distributed with parameters described by the above moments.

Given financial statements x and background knowledge $\Im$, posterior beliefs regarding transactions are normally distributed with $E\left[y^T \mid x, \Im\right] =$

$$\begin{bmatrix} 58.183 & 15.985 & 12.198 & 1.817 & 70 & 5.748 & 30 & 22.435 & 19.764 & 0.236 \end{bmatrix}$$

and $Var\left[y \mid x, \Im\right] =$

$$\begin{bmatrix} 0.338 & 0.172 & 0.167 & -0.338 & 0 & 0.164 & 0 & 0.174 & -0.007 & 0.007 \\ 0.172 & 0.482 & -0.310 & -0.172 & 0 & 0.300 & 0 & -0.128 & -0.182 & 0.182 \\ 0.167 & -0.310 & 0.477 & -0.167 & 0 & -0.135 & 0 & 0.302 & 0.175 & -0.175 \\ -0.338 & -0.172 & -0.167 & 0.338 & 0 & -0.164 & 0 & -0.174 & 0.007 & -0.007 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.164 & 0.300 & -0.135 & -0.164 & 0 & 0.445 & 0 & -0.281 & 0.145 & -0.145 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.174 & -0.128 & 0.302 & -0.174 & 0 & -0.281 & 0 & 0.455 & -0.153 & 0.153 \\ -0.007 & -0.182 & 0.175 & 0.007 & 0 & 0.145 & 0 & -0.153 & 0.328 & -0.328 \\ 0.007 & 0.182 & -0.175 & -0.007 & 0 & -0.145 & 0 & 0.153 & -0.328 & 0.328 \end{bmatrix}$$

As our intuition suggests, the posterior mean of transactions is consistent with the financial statements, $A\left(E\left[y \mid x, \Im\right]\right) = x$, and there is no residual uncertainty regarding transactions that are not in loops, sales and cost of sales are $y_5 = 70$ and $y_7 = 30$, respectively. Next, we explore accounting accruals as a source of both valuation and evaluation information.

13.11 Smooth accruals

Now, we explore valuation and evaluation roles of smooth accruals in a simple, yet dynamic setting with informed priors regarding the initial mean of cash flows.[24] Accruals smooth cash flows to summarize the information content regarding expected cash flows from the past cash flow history. This is similar in spirit to Arya et al [2002]. In addition, we show in a moral hazard setting that the foregoing accrual statistic can be combined with current cash flows and non-accounting contractible information to efficiently (subject to *LEN* model restrictions[25]) supply incentives to replacement agents via sequential spot contracts. Informed priors regarding the permanent component of cash flows facilitates performance evaluation. The *LEN* (linear exponential normal) model application is similar to Arya et al [2004]. It is not surprising that accruals can serve as statistics for valuation or evaluation, rather the striking contribution here is that the same accrual statistic can serve both purposes without loss of efficiency.

---

[24] These examples were developed from conversations with Joel Demski, John Fellingham, and Haijin Lin.

[25] See Holmstrom and Milgrom [1987], for details on the strengths and limitations of the *LEN* (linear exponential normal) model.

13.11.1 DGP

The data generating process (DGP) is as follows. Period t cash flows (excluding the agent's compensation s) includes a permanent component m_t that derives from productive capital, the agent's contribution a_t, and a stochastic error e_t.

$$cf_t = m_t + a_t + e_t$$

The permanent component (mean) is subject to stochastic shocks.

$$m_t = g\, m_{t-1} + \epsilon_t$$

where m_0 is common knowledge (strongly informed priors), g is a deterministic growth factor, and stochastic shock ϵ_t. In addition, there exists contractible, non-accounting information that is informative of the agent's action a_t with noise μ_t.

$$y_t = a_t + \mu_t$$

Variance knowledge for the errors, e, ϵ, and μ, leads to a joint normal probability assignment with mean zero and variance-covariance matrix Σ. The *DGP* is common knowledge to management and the auditor. Hence, the auditor's role is simply to assess manager's reporting compliance with the predetermined accounting system.[26]

The agent has reservation wage RW and is evaluated subject to moral hazard. The agent's action is binary $a \in \{a_H, a_L\}$, $a_H > a_L$, with personal cost $c(a), c(a_H) > c(a_L)$, and the agent's preferences for payments s and actions are *CARA* $U(s,a) = -exp\{-r[s - c(a)]\}$. Payments are linear in performance measures w_t (with weights γ_t) plus flat wage δ_t, $s_t = \delta_t + \gamma_t^T w_t$.

The valuation role of accruals is to summarize next period's unknown expected cash flow m_{t+1} based on the history of cash flows through time t (restricted recognition). The incentive-induced equilibrium agent action a_t^* is effectively known for valuation purposes. Hence, the observable cash flow history at time t is $\{cf_1 - a_1^*, cf_2 - a_2^*, \ldots, cf_t - a_t^*\}$.

13.11.2 Valuation results

For the case $\Sigma = D$ where D is a diagonal matrix comprised of $\sigma_e^2, \sigma_\epsilon^2$, and σ_μ^2 (appropriately aligned), the following *OLS* regression identifies the most efficient valuation usage of the past cash flow history.

$$\widehat{m}_t = (H^T H)^{-1} H^T z,$$

[26] Importantly, this eliminates strategic reporting considerations typically associated with equilibrium earnings management.

$$H = \begin{bmatrix} -\nu & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \nu g & -\nu & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \nu g & -\nu \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, z = \begin{bmatrix} -\nu g\, m_0 \\ cf_1 - a_1^* \\ 0 \\ cf_2 - a_2^* \\ \vdots \\ 0 \\ cf_t - a_t^* \end{bmatrix}, \text{ and } \nu = \frac{\sigma_e}{\sigma_\varepsilon}.^{27}$$

Can accruals supply a sufficient summary of the cash flow history for the cash flow mean?[28]

We utilize difference equations to establish accruals as a valuation statistic. Let $m_t = g\, m_{t-1} + \epsilon_t, \nu = \frac{\sigma_e}{\sigma_\epsilon}$, and $\phi = \frac{\sigma_e}{\sigma_\mu}$. Also, $B = \begin{bmatrix} 1+\nu^2 & \nu^2 \\ g^2 & g^2\nu^2 \end{bmatrix} = S\Lambda S^{-1}$ where

$$\Lambda = \begin{bmatrix} \frac{1+\nu^2+g^2\nu^2-\sqrt{(1+\nu^2+g^2\nu^2)^2-4g^2\nu^4}}{2} & 0 \\ 0 & \frac{1+\nu^2+g^2\nu^2+\sqrt{(1+\nu^2+g^2\nu^2)^2-4g^2\nu^4}}{2} \end{bmatrix}$$

and

$$S = \begin{bmatrix} \frac{1+\nu^2-g^2\nu^2-\sqrt{(1+\nu^2+g^2\nu^2)^2-4g^2\nu^4}}{2g^2} & \frac{1+\nu^2-g^2\nu^2+\sqrt{(1+\nu^2+g^2\nu^2)^2-4g^2\nu^4}}{2g^2} \\ 1 & 1 \end{bmatrix}.$$

Now, define the difference equations by

$$\begin{bmatrix} den_t \\ num_t \end{bmatrix} = B^t \begin{bmatrix} den_0 \\ num_0 \end{bmatrix} = S\Lambda^t S^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The primary result for accruals as a valuation statistic is presented in proposition 13.1.[29]

Proposition 13.1 *Let $m_t = g\, m_{t-1} + e_t, \Sigma = D$, and $\nu = \frac{\sigma_e}{\sigma_\epsilon}$. Then, $accruals_{t-1}$ and cf_t are, collectively, sufficient statistics for the mean of cash flows m_t based on the history of cash flows and $g^{t-1}accruals_t$ is an efficient statistic for m_t*

$$\begin{aligned} [\widehat{m}_t | cf_1, ..., cf_t] &= g^{t-1} accruals_t \\ &= \frac{1}{den_t} \left\{ \frac{num_t}{g^2}(cf_t - a_t^*) + g^{t-1}\nu^2 den_{t-1} accruals_{t-1} \right\} \end{aligned}$$

where $accruals_0 = m_0$, and $\begin{bmatrix} den_t \\ num_t \end{bmatrix} = B^t \begin{bmatrix} den_0 \\ num_0 \end{bmatrix} = S\Lambda^t S \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. The variance of accruals is equal to the variance of the estimate of the mean of cash

[27] Other information, y_t, is suppressed as it isn't informative for the cash flow mean.

[28] As the agent's equilibrium contribution a^* is known, expected cash flow for the current period is estimated by $\widehat{m}_t + a_t^*$ and next period's expected cash flow is predicted by $g\, \widehat{m}_t + a_{t+1}^*$.

[29] All proofs are included in the end of chapter appendix.

flows multiplied by $g^{2(t-1)}$; the variance of the estimate of the mean of cash flows equals the coefficient on current cash flow multiplied by σ_e^2, $Var[\widehat{m}_t] = \frac{num_t}{den_t g^2}\sigma_e^2$.

Hence, the current accrual equals the estimate of the current mean of cash flows scaled by g^{t-1}, $accruals_t = \frac{1}{g^{t-1}}\widehat{m}_t$.

Tidy accruals

To explore the tidiness property of accruals in this setting it is instructive to consider the weight placed on the most recent cash flow as the number of periods becomes large. This limiting result is expressed in corollary 13.2.

Corollary 13.2 *As t becomes large, the weight on current cash flows for the efficient estimator of the mean of cash flows approaches*

$$\frac{2}{1+(1-g^2)\nu^2+\sqrt{\left(1+(1+g^2)\nu^2\right)^2-4g^2\nu^4}}$$

and the variance of the estimate approaches

$$\frac{2}{1+(1-g^2)\nu^2+\sqrt{\left(1+(1+g^2)\nu^2\right)^2-4g^2\nu^4}}\sigma_e^2.$$

Accruals, as identified above, are tidy in the sense that each period's cash flow is ultimately recognized in accounting income or remains as a "permanent" amount on the balance sheet.[30] This permanent balance is approximately

$$\sum_{t=1}^{k-1} cf_t\left[1-\frac{num_t}{num_k}-num_t\sum_{n=t}^{k-1}\frac{g^{n-t-2}\nu^{2(n-1)}}{g^{n-1}den_n}\right]$$

where k is the first period where $\frac{num_t}{g^2 den_t}$ is well approximated by the asymptotic rate identified in corollary 1 and the estimate of expected cash flow $\widehat{m}_t$ is identified from tidy accruals as $g^{t-1}accruals_t$.[31]

In the benchmark case ($\Sigma = \sigma_e^2 I$, $\nu = \phi = 1$, and $g = 1$), this balance reduces to

$$\sum_{t=1}^{k-1} cf_t\left[1-\frac{F_{2t}}{F_{2k}}-F_{2t}\sum_{n=t}^{k-1}\frac{1}{F_{2n+1}}\right]$$

where the estimate of expected cash flow $\widehat{m}_t$ is equal to tidy $accruals_t$.

[30] The permanent balance is of course settled up on disposal or dissolution.

[31] Cash flows beginning with period k and after are fully accrued as the asymptotic rate effectively applies each period. Hence, a convergent geometric series is formed that sums to one. On the other hand, the permanent balance arises as a result of the influence of the common knowledge initial expected cash flow m_0.

13.11.3 Performance evaluation

On the other hand, the evaluation role of accruals must regard a_t as unobservable while previous actions of this or other agents are at the incentive-induced equilibrium action a^*, and all observables are potentially (conditionally) informative: $\{cf_1 - a_1^*, cf_2 - a_2^*, \ldots, cf_t\}$, and $\{y_1 - a_1^*, y_2 - a_2^*, \ldots, y_t\}$.[32]

For the case $\Sigma = D$, the most efficient linear contract can be found by determining the incentive portion of compensation via *OLS* and then plugging a constant δ to satisfy individual rationality.[33] The (linear) incentive payments are equal to the *OLS* estimator, the final element of $\widehat{a}_t$, multiplied by $\Delta = \frac{c(a_H)-c(a_L)}{a_H-a_L}$, $\gamma_t = \Delta\, \widehat{a}_t$ where[34]

$$\widehat{a}_t = (H_a^T H_a)^{-1} H_a^T w_t,$$

$$H_a = \begin{bmatrix} -\nu & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ \nu g & -\nu & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \nu g & -\nu & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 \\ 0 & 0 & \cdots & 0 & 0 & \phi \end{bmatrix}, w_t = \begin{bmatrix} -\nu g\, m_0 \\ cf_1 - a_1^* \\ 0 \\ cf_2 - a_2^* \\ \vdots \\ 0 \\ cf_t \\ \phi y_t \end{bmatrix}, \text{ and } \phi = \frac{\sigma_e}{\sigma_\varepsilon}.$$

Further, the variance of the incentive payments equals the last row, column element of $\Delta^2 (H_a^T H_a)^{-1} \sigma_e^2$.

In a moral hazard setting, the incentive portion of the *LEN* contract based on cash flow and other monitoring information history is identified in proposition 13.3. Incentive payments depend only on two realizations: unexpected cash flow and other monitoring information for period t. Unexpected cash flow at time t is

$$\begin{aligned} cf_t - E[cf_t | cf_1, \ldots, cf_{t-1}] &= cf_t - g^{t-1} accruals_{t-1} \\ &= cf_t - \widehat{m}_{t-1} \\ &= cf_t - [\widehat{m}_t | cf_1, \ldots, cf_{t-1}]. \end{aligned}$$

As a result, sequential spot contracting with replacement agents has a particularly streamlined form. Accounting accruals supply a convenient and sufficient summary of the cash flow history for the cash flow mean. Hence, the combination of last period's accruals with current cash flow yields the pivotal unexpected cash flow variable.

---

[32] For the case $\Sigma = D$, past y's are uninformative of the current period's act.

[33] Individual rationality is satisfied if

$\delta = RW - \{E[incentive payments|a] - \frac{1}{2} rVar[s] - c(a)\}$.

[34] The nuisance parameters (the initial $2t$ elements of $\widehat{a}_t$) could be avoided if one employs *GLS* in place of *OLS*.

Proposition 13.3 *Let* $m_t = g\, m_{t-1} + e_t$, $\Sigma = D$, $\nu = \frac{\sigma_e}{\sigma_\epsilon}$, *and* $\phi = \frac{\sigma_e}{\sigma_\mu}$. *Then,* $accruals_{t-1}$, cf_t, *and* y_t, *collectively, are sufficient statistics for evaluating the agent with incentive payments given by*

$$\gamma_t^T w_t = \Delta \frac{1}{\nu^2 den_{t-1} + \phi^2 den_t} \left[\begin{array}{c} \phi^2 den_t y_t \\ +\nu^2 den_{t-1} \left(cf_t - g^{t-1} accruals_{t-1} \right) \end{array} \right]$$

and variance of payments equal to

$$Var[\gamma_t^T w_t] = \Delta^2 \frac{den_t}{\nu^2 den_{t-1} + \phi^2 den_t} \sigma_e^2$$

where $\Delta = \frac{c(a_H) - c(a_L)}{aH - aL}$, *and* $accruals_{t-1}$ *and* den_t *are as defined in proposition 13.1.*

Benchmark case

Suppose $\Sigma = \sigma_e^2 I$ $(\nu = \phi = 1)$ and $g = 1$. This benchmark case highlights the key informational structure in the data. Corollary 13.4 identifies the linear combination of current cash flows and last period's accruals employed to estimate the current cash flow mean conditional on cash flow history for this benchmark case.

Corollary 13.4 *For the benchmark case* $\Sigma = \sigma_e^2 I$ $(\nu = \phi = 1)$ *and* $g = 1$, *accruals at time* t *are an efficient summary of past cash flow history for the cash flow mean if*

$$\begin{aligned} [\widehat{m}_t | cf_1, ..., cf_t] &= accruals_t \\ &= \frac{F_{2t}}{F_{2t+1}} (cf_t - a_t^*) + \frac{F_{2t-1}}{F_{2t+1}} accruals_{t-1} \end{aligned}$$

where $F_n = F_{n-1} + F_{n-2}$, $F_0 = 0$, $F_1 = 1$ *(the Fibonacci series), and the sequence is initialized with* $accruals_0 = m_0$ *(common knowledge mean beliefs). Then, variance of accruals equals* $Var\left[\widehat{m}_t\right] = \frac{F_{2t}}{F_{2t+1}} \sigma_e^2$.

For the benchmark case, the evaluation role of accruals is synthesized in corollary 13.5.

Corollary 13.5 *For the benchmark case* $\Sigma = \sigma_e^2 I$ $(\nu = \phi = 1)$ *and* $g = 1$, $accruals_{t-1}$, cf_t, *and* y_t *are, collectively, sufficient statistics for evaluating the agent with incentive payments given by*

$$\gamma_t^T w_t = \Delta \left\{ \frac{F_{2t+1}}{L_{2t}} y_t + \frac{F_{2t-1}}{L_{2t}} \left(cf_t - accruals_{t-1} \right) \right\}$$

and variance of payments equals $\Delta^2 \frac{F_{2t+1}}{L_{2t}} \sigma_e^2$ *where* $accruals_{t-1}$ *is as defined in corollary 2,* $L_n = L_{n-1} + L_{n-2}$, $L_0 = 2$, $L_1 = 1$ *(the Lucas series), and* $\Delta = \frac{c(a_H) - c(a_L)}{a_H - a_L}$.[35]

[35] The Lucas and Fibonacci series are related by $L_n = F_{n-1} + F_{n+1}$, for $n = 1, 2, \ldots$.

13.11.4 Summary

A positive view of accruals is outlined above. Accruals combined with current cash flow can serve as sufficient statistics of the cash flow history for the mean of cash flows. Further, in a moral hazard setting accruals can be combined with current cash flow and other monitoring information to efficiently evaluate replacement agents via sequential spot contracts. Informed priors regarding the contaminating permanent component facilitates this performance evaluation exercise. Notably, the same accrual statistic serves both valuation and evaluation purposes.

Next, we relax common knowledge of the *DGP* by both management and the auditor to explore strategic reporting equilibria albeit with a simpler *DGP*. That is, we revisit earnings management with informed priors and focus on Bayesian separation of signal (regarding expected cash flows) from noise.

13.12 Earnings management

We return to the earnings management setting introduced in chapter 2 and continued in chapter 3.[36] Now, we focus on belief revision with informed priors. First, we explore stochastic manipulation, as before, and, later on, selective manipulation.

13.12.1 Stochastic manipulation

The analyst is interested in uncovering the mean of accruals $E[x_t] = \mu$ (for all t) from a sequence of reports $\{y_t\}$ subject to stochastic manipulation by management. Earnings management is curbed by the auditor such that manipulation is limited to δ. That is, reported accruals y_t equal privately observed accruals x_t when there is no manipulation $I_t = 0$ and add δ when there is manipulation $I_t = 1$

$$\begin{array}{ll} y_t = x_t & \Pr(I_t = 0) = 1 - \alpha \\ y_t = x_t + \delta & \Pr(I_t = 1) = \alpha \end{array}$$

The (prior) probability of manipulation α is known as well as the variance of x_t, σ_d^2. Since the variance is known, the maximum entropy likelihood function for the data is Gaussian with unknown, but finite and constant, mean. Background knowledge regarding the mean of x_t is that the mean is μ_0 with variance σ_0^2. Hence, the maximum entropy prior distribution for the mean is also Gaussian. And, the analysts' interests focus on the mean of the posterior distribution for x, $E\left[\mu \mid \mu_0, \sigma_0^2, \sigma_d^2, \{y_t\}\right]$.

Consider the updating of beliefs when the first report is observed, y_1. The analyst knows

$$\begin{array}{ll} y_1 = x_1 & I_1 = 0 \\ y_1 = x_1 + \delta & I_1 = 1 \end{array}$$

[36] These examples were developed from conversations with Joel Demski and John Fellingham.

plus the prior probability of manipulation is α. The report contains evidence regarding the likelihood of manipulation. Thus, the posterior probability of manipulation[37] is

$$\begin{aligned} p_1 &\equiv \Pr\left(I_1 = 1 \mid \mu_0, \sigma_0^2, \sigma_d^2, y_1\right) \\ &= \frac{\alpha\phi\left(\frac{y_1-\delta-\mu_0}{\sqrt{\sigma_d^2+\sigma_0^2}}\right)}{\alpha\phi\left(\frac{y_1-\delta-\mu_0}{\sqrt{\sigma_d^2+\sigma_0^2}}\right) + (1-\alpha)\,\phi\left(\frac{y_1-\mu_0}{\sqrt{\sigma_d^2+\sigma_0^2}}\right)} \end{aligned}$$

where $\phi(\cdot)$ is the standard Normal (Gaussian) density function. The density functions are, of course, conditional on manipulation or not and the random variable of interest is $x_1 - \mu_0$ which is Normally distributed with mean zero and variance $\sigma_d^2 + \sigma_0^2 = \sigma_d^2\left(1 + \frac{1}{\nu^2}\right)$ where $\nu = \frac{\sigma_d}{\sigma_0}$.

Bayesian updating of the mean following the first report is

$$\begin{aligned} \mu_1 &= \mu_0 + \sigma_1^2 \frac{1}{\sigma_d^2}\left(p_1\left(y_1 - \delta\right) + \left(1 - p_1\right) y_1 - \mu_0\right) \\ &= \frac{1}{\nu^2+1}\left[\nu^2\mu_0 + p_1\left(y_1 - \delta\right) + \left(1 - p_1\right) y_1\right] \end{aligned}$$

where the variance of the estimated mean is

$$\begin{aligned} \sigma_1^2 &= \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_d^2}} \\ &= \frac{\sigma_d^2}{\nu^2+1} \end{aligned}$$

Since

$$Var\left[p_t\left(y_t - \delta \mid I_t = 1\right) + \left(1 - p_t\right)\left(y_t \mid I_t = 0\right)\right] = Var\left[x_t\right] \equiv \sigma_d^2 \quad \text{for all } t$$

$\sigma_1^2, \ldots, \sigma_t^2$ are known in advance of observing the reported data. That is, the information matrix is updated each period in a known way.

---

[37] The posterior probability is logistic distributed (see Kiefer [1980]).

$$p_t = \frac{1}{1 + Exp\left[a_t + b_t y_t\right]}$$

where

$$a_t = \ln\left(\frac{1-\alpha}{\alpha}\right) + \frac{1}{2\left(\sigma_d^2 + \sigma_{t-1}^2\right)}\left[\left(\delta + \mu_{t-1}\right)^2 - \mu_{t-1}^2\right]$$

and

$$b_t = \frac{1}{\left(\sigma_d^2 + \sigma_{t-1}^2\right)}\left[\mu_{t-1}^2 - \left(\delta + \mu_{t-1}\right)^2\right]$$

This updating is repeated each period.[38] The posterior probability of manipulation given the series of observed reports through period t is

$$\begin{aligned} p_t &\equiv \Pr\left(I_t = 1 \mid \mu_0, \sigma_0^2, \sigma_d^2, \{y_t\}\right) \\ &= \frac{\alpha\phi\left(\frac{y_t-\delta-\mu_{t-1}}{\sqrt{\sigma_d^2+\sigma_{t-1}^2}}\right)}{\alpha\phi\left(\frac{y_t-\delta-\mu_{t-1}}{\sqrt{\sigma_d^2+\sigma_{t-1}^2}}\right) + (1-\alpha)\,\phi\left(\frac{y_1-\mu_{t-1}}{\sqrt{\sigma_d^2+\sigma_{t-1}^2}}\right)} \end{aligned}$$

where the random variable of interest is $x_t - \mu_{t-1}$ which is Normally distributed with mean zero and variance $\sigma_d^2 + \sigma_{t-1}^2$. The updated mean is

$$\begin{aligned} \mu_t &= \mu_{t-1} + \sigma_t^2 \frac{1}{\sigma_d^2}\left(p_t\left(y_t - \delta\right) + \left(1 - p_t\right) y_t - \mu_{t-1}\right) \\ &= \frac{1}{\nu^2 + t}\left[\nu^2\mu_0 + \sum_{k=1}^{t} p_k\left(y_k - \delta\right) + \left(1 - p_k\right) y_k\right] \end{aligned}$$

and the updated variance of the mean is[39]

$$\begin{aligned} \sigma_t^2 &= \frac{1}{\frac{1}{\sigma_0^2} + t\frac{1}{\sigma_d^2}} \\ &= \frac{\sigma_d^2}{\nu^2 + t} \end{aligned}$$

---

[38] To see this as a standard conditional Gaussian distribution result, suppose there is no manipulation so that $x_1, \ldots, x_t$ are observed and we're interested in $E\left[\mu \mid x_1, \ldots, x_t\right]$ and $Var\left[\mu \mid x_1, \ldots, x_t\right]$. The conditional distribution follows immediately from the joint distribution of

$$\mu = \mu_0 + \eta_0$$

$$x_1 = \mu + \varepsilon_1 = \mu_0 + \eta_0 + \varepsilon_1$$

and so on

$$x_t = \mu + \varepsilon_t = \mu_0 + \eta_0 + \varepsilon_t$$

The joint distribution is multivariate Gaussian

$$N\left(\begin{bmatrix} \mu_0 \\ \mu_0 \\ \vdots \\ \mu_0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_0^2 & \sigma_0^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 + \sigma_d^2 & \sigma_0^2 & \sigma_0^2 \\ \sigma_0^2 & \vdots & \ddots & \vdots \\ \sigma_0^2 & \sigma_0^2 & \cdots & \sigma_0^2 + \sigma_d^2 \end{bmatrix}\right)$$

With manipulation, the only change is x_t is replaced by $(y_t - \delta \mid I_t = 1)$ with probability p_t and $(y_t \mid I_t = 0)$ with probability $1 - p_t$.

[39] Bayesian updating of the mean can be thought of as a stacked weighted projection exercise where the prior "sample" is followed by the new evidence. For period t, the updated mean is

$$\mu_t \equiv E\left[\mu \mid \mu_0, \{y_t\}\right] = \left(X_t^T X_t\right)^{-1} X_t^T Y_t$$

and the updated variance of the mean is

$$\sigma_t^2 \equiv Var\left[\mu \mid \mu_0, \{y_t\}\right] = \left(X_t^T X_t\right)^{-1}$$

Now, it's time to look at some data.

Experiment

Suppose the prior distribution for x has mean $\mu_0 = 100$ and standard deviation $\sigma_0 = 25$, then it follows (from maximum entropy) the prior distribution is Gaussian. Similarly, x_t is randomly sampled from a Gaussian distribution with mean μ, where the value of μ is determined by a random draw from the prior distribution $N(100, 25)$, and standard deviation $\sigma_d = 20$. Reports y_t are stochastically manipulated as $x_t + \delta$ with likelihood $\alpha = 0.2$, where $\delta = 20$, and $y_t = x_t$ otherwise.

Results

Two plots summarize the data. The first data plot, figure 13.4, depicts the mean of 100 simulated samples of $t = 100$ observations and the mean of the 95% interval estimates of the mean along with the baseline (dashed line) for the randomly drawn mean μ of the data. As expected, the mean estimates converge toward the baseline as t increases and the interval estimates narrow around the baseline.

The second data plot, figure 13.5, shows the incidence of manipulation along with the assessed posterior probability of manipulation (multiplied by δ) based on the report for a representative draw. The graph depicts a reasonably tight correspondence between incidence of manipulation and posterior beliefs regarding manipulation.

Scale uncertainty

Now, we consider a setting where the variance (scale parameter) associated with privately observed accruals, σ_d^2, and the prior, σ_0^2, are uncertain. Suppose we only

---

where

$$Y_t = \begin{bmatrix} \frac{1}{\sigma_0}\mu_0 \\ \frac{\sqrt{p_1}}{\sigma_d}(y_1 - \delta) \\ \frac{\sqrt{1-p_1}}{\sigma_d} y_1 \\ \vdots \\ \frac{\sqrt{p_t}}{\sigma_d}(y_t - \delta) \\ \frac{\sqrt{1-p_t}}{\sigma_d} y_t \end{bmatrix}$$

and

$$X_t = \begin{bmatrix} \frac{1}{\sigma_0} \\ \frac{\sqrt{p_1}}{\sigma_d} \\ \frac{\sqrt{1-p_1}}{\sigma_d} \\ \vdots \\ \frac{\sqrt{p_t}}{\sigma_d} \\ \frac{\sqrt{1-p_t}}{\sigma_d} \end{bmatrix}$$

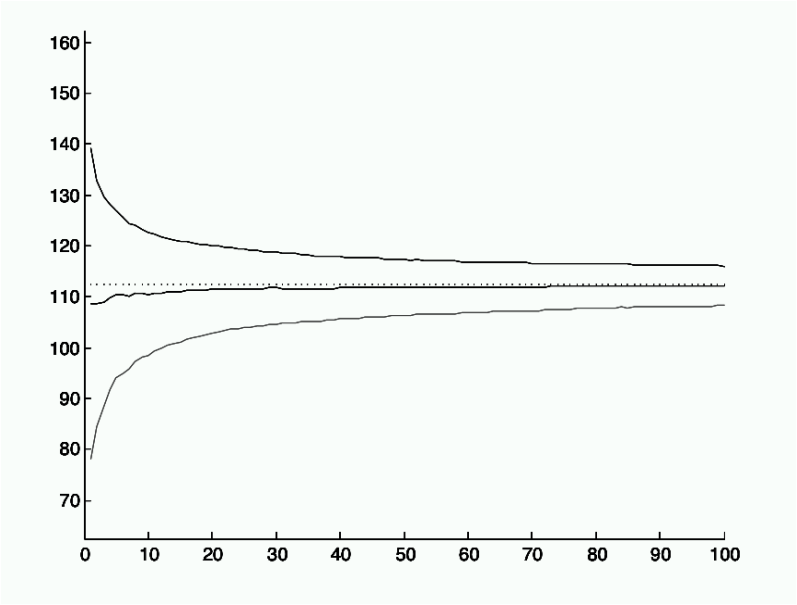

Figure 13.4: Stochastic manipulation σ_d known

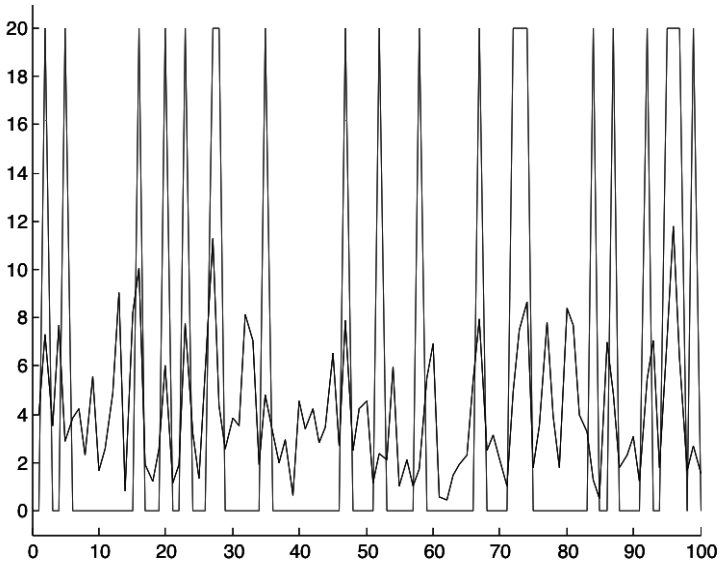

Figure 13.5: Incidence of stochastic manipulation and posterior probability

know $\nu = \frac{\sigma_d}{\sigma_0}$ and σ_d^2 and σ_0^2 are positive. Then, Jeffreys' prior distribution for scale is proportional to the square root of the determinant of the information matrix for t reports $\{y_t\}$ (see Jaynes [2003]),

$$f(\sigma_d) \propto \sqrt{\frac{\nu^2 + t}{\sigma_d^2}}$$

Hence, the prior for scale is proportional to $\frac{1}{\sigma_d}$

$$f(\sigma_d) \propto \frac{1}{\sigma_d}$$

With the mean μ and scale σ_d^2 unknown, following Box and Tiao [1973, p. 51], we can write the likelihood function (with priors on the mean incorporated as above) as

$$\ell\left(\mu, \sigma_d^2 \mid \{y_t\}\right) = \left(\sigma_d^2\right)^{-\frac{t+1}{2}} \exp\left[-\frac{1}{2\sigma_d^2}(Y - X\mu)^T (Y - X\mu)\right]$$

Now, rewrite[40]

$$\begin{aligned}(Y - X\mu)^T (Y - X\mu) &= \left(Y - \widehat{Y}\right)^T \left(Y - \widehat{Y}\right) \\ &\quad + \left(\widehat{Y} - X\mu\right)^T \left(\widehat{Y} - X\mu\right) \\ &= ts_t^2 + (\mu - \mu_t)^T X^T X (\mu - \mu_t)\end{aligned}$$

---

[40]The decomposition is similar to decomposition of mean square error into variance and squared bias but without expectations. Expand both sides of

$$(Y - X\mu)^T (Y - X\mu) = \left(Y - \widehat{Y}\right)^T \left(Y - \widehat{Y}\right) + \left(\widehat{Y} - X\mu\right)^T \left(\widehat{Y} - X\mu\right)$$

The left hand side is

$$Y^T Y - 2Y^T X\mu + \mu^T X^T X\mu$$

The right hand side is

$$Y^T Y - 2Y^T X\widehat{\mu} + \widehat{\mu}^T X^T X\widehat{\mu} + \mu^T X^T X\mu - 2\mu^T X^T X\widehat{\mu} + \widehat{\mu}^T X^T X\widehat{\mu}$$

Now show

$$-2Y^T X\mu = -2Y^T X\widehat{\mu} + 2\widehat{\mu}^T X^T X\widehat{\mu} - 2\widehat{\mu}^T X^T X\mu$$

Rewriting yields

$$Y^T X(\widehat{\mu} - \mu) = \widehat{\mu}^T X^T X(\widehat{\mu} - \mu)$$

or combining

$$\begin{aligned}(Y - X\widehat{\mu})^T X(\widehat{\mu} - \mu) &= 0 \\ \widehat{\varepsilon}^T X(\widehat{\mu} - \mu) &= 0\end{aligned}$$

The last expression is confirmed as $X^T\widehat{\varepsilon} = 0$ by least squares estimator construction (the residuals $\widehat{\varepsilon}$ are chosen to be orthogonal to the columns of X).

where

$$s_t^2 = \frac{1}{t}\left(Y - \widehat{Y}\right)^T \left(Y - \widehat{Y}\right)$$

$$Y^T = \left[\begin{array}{cccccc} \nu\mu_0 & \sqrt{p_1}y_1 & \sqrt{1-p_1}y_1 & \cdots & \sqrt{p_t}y_t & \sqrt{1-p_t}y_t \end{array}\right]$$

$$\widehat{Y} = X\mu_t$$

$$X^T = \left[\begin{array}{cccccc} \nu & \sqrt{p_1} & \sqrt{1-p_1} & \cdots & \sqrt{p_t} & \sqrt{1-p_t} \end{array}\right]$$

$$\mu_t = \left(X^T X\right)^{-1} X^T Y$$

Hence,

$$\begin{aligned} \ell\left(\mu, \sigma_d^2 \mid \{y_t\}\right) &= \left(\sigma_d^2\right)^{-\frac{t+1}{2}} \exp\left[-\frac{ts_t^2}{2\sigma_d^2} - \frac{\left(\mu-\mu_t\right)^T X^T X\left(\mu-\mu_t\right)}{2\sigma_d^2}\right] \\ &= \left(\sigma_d^2\right)^{-\frac{t+1}{2}} \exp\left[-\frac{ts_t^2}{2\sigma_d^2}\right] \exp\left[-\frac{\left(\mu-\mu_t\right)^T X^T X\left(\mu-\mu_t\right)}{2\sigma_d^2}\right] \end{aligned}$$

The posterior distribution for the unknown parameters is then

$$f\left(\mu, \sigma_d^2 \mid \{y_t\}\right) \propto \ell\left(\mu, \sigma_d^2 \mid \{y_t\}\right) f\left(\sigma_d\right)$$

substitution from above gives

$$\begin{aligned} f\left(\mu, \sigma_d^2 \mid \{y_t\}\right) \propto\ & \left(\sigma_d^2\right)^{-\left(\frac{t}{2}+1\right)} \exp\left[-\frac{ts_t^2}{2\sigma_d^2}\right] \\ & \times \exp\left[-\frac{\left(\mu-\mu_t\right)^T X^T X\left(\mu-\mu_t\right)}{2\sigma_d^2}\right] \end{aligned}$$

The posterior decomposes into

$$f\left(\mu, \sigma_d^2 \mid \{y_t\}\right) = f\left(\sigma_d^2 \mid s_t^2\right) f\left(\mu \mid \mu_t, \sigma_d^2\right)$$

where

$$f\left(\mu \mid \mu_t, \sigma_d^2\right) \propto \exp\left[-\frac{\left(\mu-\mu_t\right)^T X^T X\left(\mu-\mu_t\right)}{2\sigma_d^2}\right]$$

is the multivariate Gaussian kernel, and

$$f\left(\sigma_d^2 \mid s_t^2\right) \propto \left(\sigma_d^2\right)^{-\left(\frac{t}{2}+1\right)} \exp\left[-\frac{ts_t^2}{2\sigma_d^2}\right],\ t \geq 1$$

is the inverted chi-square kernel, which is conjugate prior to the variance of a Gaussian distribution. Integrating out σ_d^2 yields the marginal posterior for μ,

$$f\left(\mu \mid \{y_t\}\right) = \int_0^\infty f\left(\mu, \sigma_d^2 \mid \{y_t\}\right) d\sigma_d^2$$

which has a noncentral, scaled-Student $\mathrm{t}\left(\mu_t, s_t^2 \left(X^T X\right)^{-1}, t\right)$ distribution. In other words,

$$T = \frac{\mu - \mu_t}{\frac{s_t}{\sqrt{\nu^2 + t}}}$$

has a Student $\mathrm{t}(t)$ distribution, for $t \geq 1$ (see Box and Tiao [1973, p. 117-118]).[41]

Now, the estimate for μ conditional on reports to date is the posterior mean[42]

$$\begin{aligned} \mu_t &\equiv E\left[\mu \mid \{y_t\}\right] \\ &= \frac{\int_{-\infty}^{\infty} \mu f\left(\mu \mid \{y_t\}\right) d\mu}{\int_{-\infty}^{\infty} f\left(\mu \mid \{y_t\}\right) d\mu} \\ &= \frac{\nu^2 \mu_0 + p_1 y_1 + (1 - p_1) y_1 + \cdots + p_t y_t + (1 - p_t) y_t}{\nu^2 + t} \end{aligned}$$

from the above posterior distribution and p_t is defined below. The variance of the estimate for μ is

$$\begin{aligned} \sigma_t^2 &\equiv Var\left[\mu_t\right] \\ &= \widetilde{s}_t^2 \left(X^T X\right)^{-1} = \frac{\widetilde{s}_t^2}{t + \nu^2},\ t \geq 1 \end{aligned}$$

where $\widetilde{s}_t^2$ is the estimated variance of the posterior distribution for x_t (see discussion below under a closer look at the variance). Hence, the highest posterior density (most compact) interval for μ with probability p is

$$\mu_t \pm \mathrm{t}\left(t; 1 - \frac{p}{2}\right) \sigma_t$$

$$\begin{array}{c} \frac{\nu^2 \mu_0 + p_1 y_1 + (1 - p_1) y_1 + \cdots + p_t y_t + (1 - p_t) y_t}{\nu^2 + t} \\ \pm \mathrm{t}\left(t; 1 - \frac{p}{2}\right) \frac{\widetilde{s}_t}{\sqrt{t + \nu^2}} \end{array} \qquad t \geq 1$$

---

[41] This follows from a transformation of variables,

$$z = \frac{A}{2\sigma_d^2}$$

where

$$A = t s^2 + (\mu - \mu_t)^T X^T X (\mu - \mu_t)$$

that produces the kernel of a scaled Student t times the integral of a gamma distribution (see Gelman et al [2004], p.76). Or, for $a > 0, p > 0$,

$$\int_0^{\infty} x^{-(p+1)} e^{-\frac{a}{x}} dx = a^{-p} \Gamma(p)$$

where

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$$

and for n a positive integer

$$\Gamma(n) = (n - 1)!$$

a constant which can be ignored when identifying the marginal posterior (see Box and Tiao [1973, p. 144]).

[42] For emphasis, we write the normalization factor in the denominator of the expectations expression.

The probability the current report y_t, for $t \geq 2$,[43] is manipulated conditional on the history of reports (and manipulation probabilities) is

$$\begin{aligned} p_t &\equiv \Pr\left(I_t = 1 \mid \nu, \{y_t\}, \{p_{t-1}\}, \mu_{t-1}, \sigma^2_{t-1}\right),\ t \geq 2 \\ &= \frac{\alpha \int\int f\left(y_t | D_t = 1, \mu, \sigma^2_d\right) f\left(\mu | \mu_{t-1}, \sigma^2_d\right) f\left(\sigma^2_d | s^2_{t-1}\right) d\mu d\sigma^2_d}{den(p_t)} \\ = \frac{\alpha \int\int \left(\sigma^2_d\right)^{-\frac{1}{2}} \exp\left[-\frac{(y_t-\delta-\mu)^2}{2\sigma^2_d}\right] \left(\sigma^2_d\right)^{-\frac{1}{2}} \exp\left[-\frac{(\mu-\mu_{t-1})X^T X(\mu-\mu_{t-1})}{2\sigma^2_d}\right] d\mu}{den(p_t)} & \\ \times \left(\sigma^2_d\right)^{-\left(\frac{t}{2}+1\right)} \exp\left[-\frac{(t-1)s^2_{t-1}}{2\sigma^2_d}\right] d\sigma^2_d & \\ + (1-\alpha) \int \frac{1}{\sqrt{\sigma^2_d + \sigma^2_{t-1}}} \exp\left[-\frac{1}{2}\frac{\left(y_t - \mu_{t-1}\right)^2}{\sigma^2_d + \sigma^2_{t-1}}\right] & \\ \times \left(\sigma^2_d\right)^{-\left(\frac{t}{2}+1\right)} \exp\left[-\frac{(t-1)s^2_{t-1}}{2\sigma^2_d}\right] d\sigma^2_d & \end{aligned}$$

where

$$\begin{aligned} den\left(p_t\right) &= \alpha \int \frac{1}{\sqrt{\sigma^2_d + \sigma^2_{t-1}}} \exp\left[-\frac{1}{2}\frac{\left(y_t - \delta - \mu_{t-1}\right)^2}{\sigma^2_d + \sigma^2_{t-1}}\right] \\ &\quad \times \left(\sigma^2_d\right)^{-\left(\frac{t}{2}+1\right)} \exp\left[-\frac{(t-1)\, s^2_{t-1}}{2\sigma^2_d}\right] d\sigma^2_d \\ &\quad + (1-\alpha) \int \frac{1}{\sqrt{\sigma^2_d + \sigma^2_{t-1}}} \exp\left[-\frac{1}{2}\frac{\left(y_t - \mu_{t-1}\right)^2}{\sigma^2_d + \sigma^2_{t-1}}\right] \\ &\quad \times \left(\sigma^2_d\right)^{-\left(\frac{t}{2}+1\right)} \exp\left[-\frac{(t-1)\, s^2_{t-1}}{2\sigma^2_d}\right] d\sigma^2_d \end{aligned}$$

Now, we have

$$\begin{aligned} f\left(y_t - \delta \mid D_t = 1\right) = &\int_0^\infty \frac{1}{\sqrt{\sigma^2_d + \sigma^2_{t-1}}} \exp\left[-\frac{1}{2}\frac{\left(y_t - \delta - \mu_{t-1}\right)^2}{\sigma^2_d + \sigma^2_{t-1}}\right] \\ &\times \left(\sigma^2_d\right)^{-\left(\frac{t}{2}+1\right)} \exp\left[-\frac{(t-1)s^2_{t-1}}{2\sigma^2_d}\right] d\sigma^2_d \end{aligned}$$

and

$$\begin{aligned} f\left(y_t \mid D_t = 0\right) &= \int_0^\infty \frac{1}{\sqrt{\sigma^2_d + \sigma^2_{t-1}}} \exp\left[-\frac{1}{2}\frac{\left(y_t - \mu_{t-1}\right)^2}{\sigma^2_d + \sigma^2_{t-1}}\right] \\ &\quad \times \left(\sigma^2_d\right)^{-\left(\frac{t}{2}+1\right)} \exp\left[-\frac{(t-1)\, s^2_{t-1}}{2\sigma^2_d}\right] d\sigma^2_d \end{aligned}$$

[43] For $t = 1$, $p_t \equiv \Pr\left(D_t = 1 \mid y_t\right) = \alpha$ as the distribution for $\left(y_t \mid D_t\right)$ is so diffuse (s^2_0 has zero degrees of freedom) the report y_t is uninformative.

are noncentral, scaled-Student t$\left(\mu_{t-1}, s_{t-1}^2 + s_{t-1}^2 \left(X^T X\right)^{-1}, t-1\right)$ distributed. In other words,

$$T = \frac{y_t - \mu_{t-1}}{\sqrt{s_{t-1}^2 + \frac{s_{t-1}^2}{\nu^2+t-1}}}$$

has a Student t$(t-1)$ distribution for $t \geq 2$.

A closer look at the variance.

Now, we more carefully explore what s^2 estimates. We're interested in estimates of μ and $\frac{\sigma_d^2}{\nu^2+t}$ and we have the following relations:

$$\begin{aligned} x_t &= \mu + \varepsilon_t \\ &= y_t - \delta D_t \end{aligned}$$

If D_t is observed then x_t is effectively observed and estimates of $\mu = E[x]$ and $\sigma_d^2 = Var[x]$ are, by standard methods, $\overline{x}$ and s^2. However, when manipulation D_t is not observed, estimation is more subtle. $x_t = y_t - \delta D_t$ is estimated via $y_t - \delta p_t$ which deviates from x_t by $\eta_t = -\delta(D_t - p_t)$. That is, $x_t = y_t - \delta p_t + \eta_t$. where

$$E[\eta_t \mid y_t] = -\delta[p_t(1-p_t) + (1-p_t)(0-p_t)] = 0$$

and

$$\begin{aligned} Var[\eta_t \mid y_t] &= \delta^2\left[p_t(1-p_t)^2 + (1-p_t)(0-p_t)^2\right] \\ &= \delta^2 p_t(1-p_t) = \delta^2 Var[D_t \mid y_t] \end{aligned}$$

s^2 estimates $E\left[\widehat{\varepsilon}_t^T \widehat{\varepsilon}_t \mid y_t\right]$ where $\widehat{\varepsilon}_t = y_t - \delta p_t - \mu_t$. However, $\sigma_d^2 = E\left[\varepsilon_t^T \varepsilon_t\right]$ is the object of interest. We can write

$$\begin{aligned} \widehat{\varepsilon}_t &= y_t - \delta p_t - \mu_t \\ &= (\delta D_t + \mu + \varepsilon_t) - \delta p_t - \mu_t \\ &= \varepsilon_t + \delta(D_t - p_t) + (\mu - \mu_t) \end{aligned}$$

In other words,

$$\varepsilon_t + (\mu - \mu_t) = \widehat{\varepsilon}_t - \delta(D_t - p_t)$$

Since $E\left[X^T \varepsilon_t\right] = 0$ (the regression condition) and μ_t is a linear combination of X, $Cov[\varepsilon_t, (\mu - \mu_t)] = 0$. Then, the variance of the left-hand side is a function of σ_d^2, the parameter of interest.

$$\begin{aligned} Var[\varepsilon_t + (\mu - \mu_t \mid y_t)] &= Var[\varepsilon_t] + Var[\mu - \mu_t \mid y_t] \\ &= \sigma_d^2 + \sigma_d^2 \frac{1}{\nu^2 + t} \\ &= \frac{\nu^2 + t + 1}{\nu^2 + t} \sigma_d^2 \end{aligned}$$

As D_t is stochastic

$$E\left[\widehat{\varepsilon}_t\left(D_t - p_t\right) \mid y_t\right] = 0$$

the variance of the right-hand side is

$$\begin{aligned} Var\left[\widehat{\varepsilon}_t - \delta\left(D_t - p_t\right) \mid y_t\right] &= Var\left[\widehat{\varepsilon}_t \mid y_t\right] + \delta^2 Var\left[\left(D_t - p_t\right) \mid y_t\right] \\ &\quad -2\delta Cov\left[\widehat{\varepsilon}_t, \left(D_t - p_t\right) \mid y_t\right] \\ &= Var\left[\widehat{\varepsilon}_t \mid y_t\right] \\ &\quad +\delta^2\left[p_t\left(1-p_t\right)^2 + \left(1-p_t\right)\left(0-p_t\right)^2\right] \\ &= Var\left[\widehat{\varepsilon}_t \mid y_t\right] + \delta^2 p_t\left(1-p_t\right) \end{aligned}$$

As $Var\left[\widehat{\varepsilon}_t\right]$ is consistently estimated via s^2, we can estimate σ_d^2 by

$$\begin{aligned} \widehat{\sigma}_d^2 &= \frac{\nu^2 + t}{\nu^2 + t + 1}\left(s^2 + \delta^2 p_t\left(1-p_t\right)\right) \\ \widehat{\sigma}_d^2 &= \frac{\nu^2 + t}{\nu^2 + t + 1}\widetilde{s}^2 \end{aligned}$$

where $p_t\left(1-p_t\right)$ is the variance of D_t and $\widetilde{s}^2 = s^2 + \delta^2 p_t\left(1-p_t\right)$ estimates the variance of $\widehat{\varepsilon}_t + \eta_t$ given the data $\{y_t\}$.

Experiment

Repeat the experiment above except now we account for variance uncertainty as described above.[44]

Results

For 100 simulated samples of $t = 100$, we generate a plot, figure 13.6, of the mean and average 95% interval estimates. As expected, the mean estimates converge toward the baseline (dashed line) as t increases and the interval estimates narrow around the baseline but not as rapidly as the known variance setting.

---

[44] Another (complementary) inference approach involves creating the posterior distribution via conditional posterior simulation. Continue working with prior $p\left(\sigma_d^2 \mid X\right) \propto \frac{1}{\sigma_d^2}$ to generate a posterior distribution for the variance

$$p\left(\sigma_d^2 \mid X, \{y_t\}\right) \sim Inv - \chi^2\left(t, \widehat{\sigma}_d^2\right)$$

and conditional posterior distribution for the mean

$$p\left(\mu \mid \sigma_d^2, X, \{y_t\}\right) \sim N\left(\left(X^T X\right)^{-1} X^T Y, \sigma_d^2\left(X^T X\right)^{-1}\right)$$

That is, draw σ_d^2 from the inverted, scaled chi-square distribution with t degrees of freedom and scale parameter $\widehat{\sigma}_d^2$. Then draw μ from a Gaussian distribution with mean $\left(X^T X\right)^{-1} X^T Y$ and variance equal to the draw for $\sigma_d^2\left(X^T X\right)^{-1}$ from the step above.

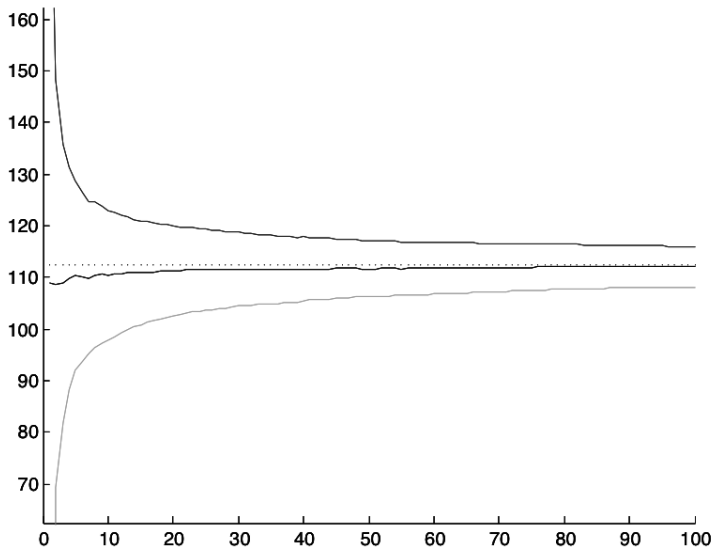

Figure 13.6: Stochastic manipulation σ_d unknown

13.12.2 Selective earnings management

Suppose earnings are manipulated whenever privately observed accruals x_t lie below prior periods' average reported accruals $\overline{y}_{t-1}$. That is,

$$\begin{array}{ll} x_t < \overline{y}_{t-1} & I_t = 1 \\ \text{otherwise} & I_t = 0 \end{array}$$

where $\overline{y}_0 = \mu_0$; for simplicity, μ_0 and $\overline{y}_t$ are commonly observed.[45] The setting differs from stochastic earnings management only in the prior and posterior probabilities of manipulation. The prior probability of manipulation is

$$\begin{aligned} \alpha_t &\equiv \Pr\left(x_t < \overline{y}_{t-1} \mid \mu_0, \sigma_0^2, \sigma_d^2, \{y_{t-1}\}\right) \\ &= \Phi\left(\frac{\overline{y}_{t-1} - \mu_{t-1}}{\sqrt{\sigma_d^2 + \sigma_{t-1}^2}}\right) \end{aligned}$$

where $\Phi(\cdot)$ represents the cumulative distribution function for the standard normal. Updated beliefs are informed by reported results even though they may be manipulated. If reported results exceed average reported results plus δ, then we

[45] This assumption could be relaxed or, for example, interpreted as an unbiased forecast conveyed via the firm's prospectus.

know there is no manipulation. Or, if reported results are less than average reported results less δ, then we know there is certain manipulation. Otherwise, there exists a possibility the reported results are manipulated or not. Therefore, the posterior probability of manipulation is

$$\begin{aligned} p_t \equiv & \Pr\left(I_t = 1 \mid \mu_0, \sigma_0^2, \sigma_d^2, \{y_t\}, y_t > \overline{y}_{t-1} + \delta\right) = 0 \\ & \Pr\left(I_t = 1 \mid \mu_0, \sigma_0^2, \sigma_d^2, \{y_t\}, y_t < \overline{y}_{t-1} - \delta\right) = 1 \\ & \Pr\left(I_t = 1 \mid \mu_0, \sigma_0^2, \sigma_d^2, \{y_t\}, \overline{y}_{t-1} - \delta \leq y_t \leq \overline{y}_{t-1} + \delta\right) \\ & = \frac{\alpha_t \phi\left(\frac{y_t - \delta - \mu_{t-1}}{\sqrt{\sigma_d^2 + \sigma_{t-1}^2}}\right)}{\alpha_t \phi\left(\frac{y_t - \delta - \mu_{t-1}}{\sqrt{\sigma_d^2 + \sigma_{t-1}^2}}\right) + (1 - \alpha_t)\phi\left(\frac{y_1 - \mu_{t-1}}{\sqrt{\sigma_d^2 + \sigma_{t-1}^2}}\right)} \end{aligned}$$

As before, the updated mean is

$$\begin{aligned} \mu_t &= \mu_{t-1} + \sigma_t^2 \frac{1}{\sigma_d^2}\left(p_t\left(y_t - \delta\right) + \left(1 - p_t\right) y_t - \mu_{t-1}\right) \\ &= \frac{1}{\nu^2 + t}\left[\nu^2 \mu_0 + \sum_{k=1}^{t} p_k\left(y_k - \delta\right) + \left(1 - p_k\right) y_k\right] \end{aligned}$$

and the updated variance of the mean is

$$\begin{aligned} \sigma_t^2 &= \frac{1}{\frac{1}{\sigma_0^2} + t\frac{1}{\sigma_d^2}} \\ &= \frac{\sigma_d^2}{\nu^2 + t} \end{aligned}$$

Time for another experiment.

Experiment

Suppose the prior distribution for x has mean $\mu_0 = 100$ and standard deviation $\sigma_0 = 25$, then it follows (from maximum entropy) the prior distribution is Gaussian. Similarly, x_t is randomly sampled from a Gaussian distribution with mean μ, a random draw from the prior distribution $N\left(100, 25\right)$, and standard deviation $\sigma_d = 20$. Reports y_t are selectively manipulated as $x_t + \delta$ when $x_t < \overline{y}_{t-1}$, where $\delta = 20$, and $y_t = x_t$ otherwise.

Results

Again, two plots summarize the data. The first data plot, figure 13.7, depicts the mean and average 95% interval estimates based on 100 simulated samples of $t = 100$ observations along with the baseline (dashed line) for the randomly drawn mean μ of the data. As expected, the mean estimates converge toward the baseline as t increases and the interval estimates narrow around the baseline. The second data plot, figure 13.8, shows the incidence of manipulation along with the assessed posterior probability of manipulation (multiplied by δ) based on the report for a representative draw. The graph depicts a reasonably tight correspondence between incidence of manipulation and posterior beliefs regarding manipulation.

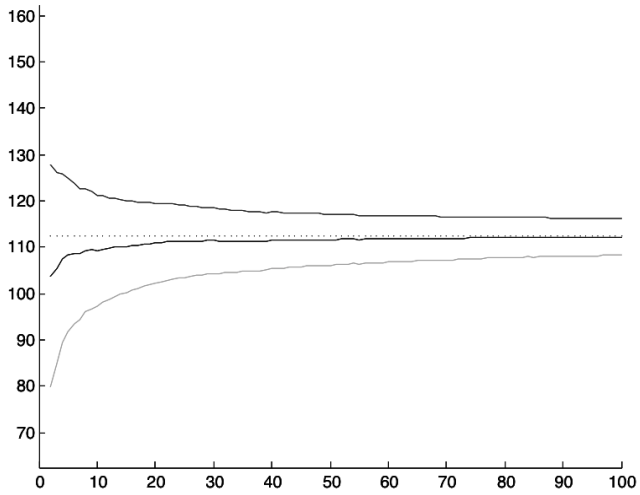

Figure 13.7: Selective manipulation σ_d known

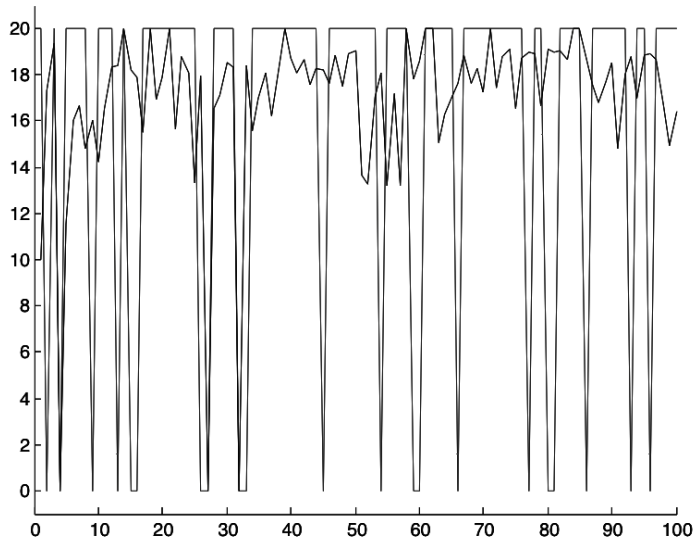

Figure 13.8: Incidence of selective manipulation and posterior probability

Scale uncertainty

Again, we consider a setting where the variance (scale parameter) associated with privately observed accruals σ_d^2 is uncertain but manipulation is selective. The only changes from the stochastic manipulation setting with uncertain scale involve the probabilities of manipulation.

The prior probability of manipulation is

$$\begin{aligned}\alpha_t &\equiv \int \Pr\left(x_t < \overline{y}_{t-1} \mid \mu_0, \nu, \sigma_d^2, \{y_{t-1}\}\right) f\left(\sigma_d^2 \mid s_{t-1}^2\right) d\sigma_d^2 \\ &= \int_0^\infty \int_{-\infty}^{\overline{y}_{t-1}} f\left(x_t \mid \mu_0, \nu, \sigma_d^2, \{y_{t-1}\}\right) dx_t f\left(\sigma_d^2 \mid s_{t-1}^2\right) d\sigma_d^2, t \geq 2\end{aligned}$$

On integrating σ_d^2 out, the prior probability of manipulation then simplifies as

$$\alpha_t = \int_{-\infty}^{\overline{y}_{t-1}} f\left(x_t \mid \mu_0, \nu, \{y_{t-1}\}\right) dx_t, t \geq 2$$

a cumulative noncentral, scaled-Student $\mathrm{t}\left(\mu_{t-1}, s_{t-1}^2 + s_{t-1}^2 \left(X^T X\right)^{-1}, t-1\right)$ distribution; in other words,

$$T = \frac{x_t - \mu_{t-1}}{\sqrt{s_{t-1}^2 + \frac{s_{t-1}^2}{\nu^2 + t - 1}}}$$

has a Student $\mathrm{t}(t-1)$ distribution, $t \geq 2$.[46]

Following the report y_t, the posterior probability of manipulation is

$$\begin{aligned}p_t \equiv \quad & \Pr\left(I_t = 1 \mid \mu_0, \nu, \{y_t\}, y_t > \overline{y}_{t-1} + \delta\right) = 0 \\ & \Pr\left(I_t = 1 \mid \mu_0, \nu, \{y_t\}, y_t < \overline{y}_{t-1} - \delta\right) = 1 \\ & \Pr\left(I_t = 1 \mid \mu_0, \nu, \{y_t\}, \overline{y}_{t-1} - \delta \leq y_t \leq \overline{y}_{t-1} + \delta\right) \\ & = \frac{\alpha_t \int f\left(y_t \mid I_t = 1, \Im_{t-1}, \sigma_d^2\right) f\left(\sigma_d^2 \mid s_{t-1}^2\right) d\sigma_d^2}{\int f\left(y_t \mid \Im_{t-1}, \sigma_d^2\right) f\left(\sigma_d^2 \mid s_{t-1}^2\right) d\sigma_d^2}, t \geq 2\end{aligned}$$

where $\Im_{t-1} = [\mu_0, \nu, \{y_{t-1}\}]$,

$$\begin{aligned}f\left(y_t \mid \Im_{t-1}, \sigma_d^2\right) &= \alpha_t f\left(y_t - \delta \mid I_t = 1, \Im_{t-1}, \sigma_d^2\right) \\ &\quad + (1 - \alpha_t) f\left(y_t \mid I_t = 0, \Im_{t-1}, \sigma_d^2\right)\end{aligned}$$

$f\left(y_t - \delta \mid I_t = 1, \Im_{t-1}, \sigma_d^2\right)$ and $f\left(y_t \mid I_t = 0, \Im_{t-1}, \sigma_d^2\right)$ are noncentral, scaled-Student $\mathrm{t}\left(\mu_{t-1}, s_{t-1}^2 + s_{t-1}^2 \left(X^T X\right)^{-1}, t-1\right)$ distributed. In other words,

$$T = \frac{y_t - \mu_{t-1}}{\sqrt{s_{t-1}^2 + \frac{s_{t-1}^2}{\nu^2 + t - 1}}}$$

has a Student $\mathrm{t}(t-1)$ distribution for $t \geq 2$.

[46] The prior probability of manipulation is uninformed or $p_t = \frac{1}{2}$ for $t < 2$.

A closer look at the variance.

In the selective manipulation setting,

$$\begin{aligned} Var\left[\widehat{\varepsilon}_t - \delta\left(D_t - p_t\right) \mid y_t\right] &= Var\left[\widehat{\varepsilon}_t \mid y_t\right] + \delta^2 Var\left[\left(D_t - p_t\right) \mid y_t\right] \\ &\quad -2\delta E\left[\widehat{\varepsilon}_t\left(D_t - p_t\right) \mid y_t\right] \\ &= Var\left[\widehat{\varepsilon}_t \mid y_t\right] + \delta^2 p_t\left(1 - p_t\right) \\ &\quad -2\delta E\left[\widehat{\varepsilon}_t\left(D_t - p_t\right) \mid y_t\right] \end{aligned}$$

The last term differs from the stochastic setting as selective manipulation produces truncated expectations. That is,

$$\begin{aligned} 2\delta E\left[\widehat{\varepsilon}_t\left(D_t - p_t\right) \mid y_t\right] &= 2\delta\{p_t E\left[\widehat{\varepsilon}_t\left(1 - p_t\right) \mid y_t, D_t = 1\right] \\ &\quad + \left(1 - p_t\right) E\left[\widehat{\varepsilon}_t\left(0 - p_t\right) \mid y_t, D_t = 0\right]\} \end{aligned}$$

$$\begin{aligned} &= 2\delta\{p_t E\left[\widehat{\varepsilon}_t\left(1 - p_t\right) \mid y_t, x_t < \overline{y}_{t-1}\right] \\ &\quad + \left(1 - p_t\right) E\left[\widehat{\varepsilon}_t\left(0 - p_t\right) \mid y_t, x_t > \overline{y}_{t-1}\right]\} \\ &= 2\delta\{p_t E\left[\widehat{\varepsilon}_t\left(1 - p_t\right) \mid y_t, \widehat{\varepsilon}_t + \eta_t < \overline{y}_{t-1} - \mu_t\right] \\ &\quad + \left(1 - p_t\right) E\left[\widehat{\varepsilon}_t\left(0 - p_t\right) \mid y_t, \widehat{\varepsilon}_t + \eta_t > \overline{y}_{t-1} - \mu_t\right]\} \\ &= 2\delta\left\{p_t E\left[\widehat{\varepsilon}_t \mid y_t, \widehat{\varepsilon}_t + \eta_t < \overline{y}_{t-1} - \mu_t\right] - E\left[\widehat{\varepsilon}_t p_t \mid y_t\right]\right\} \\ &= 2\delta\left\{p_t \int\int \sigma\phi\left(\frac{\overline{y}_{t-1} - \mu}{\sigma} \mid \mu, \sigma\right) f\left(\mu, \sigma\right) d\mu d\sigma - 0\right\} \\ &= -2\delta p_t \widetilde{s} f\left(\frac{\overline{y}_{t-1} - \mu_t}{\widetilde{s}}\right) \end{aligned}$$

where $\widetilde{s}^2 = s^2 + \delta^2 p_t\left(1 - p_t\right)$ estimates the variance of $\widehat{\varepsilon}_t + \eta_t$ with no truncation, σ^2. The extra term, $\widetilde{s} f\left(\frac{\overline{y}_{t-1} - \mu_t}{\widetilde{s}}\right)$, arises from truncated expectations induced by selective manipulation rather than random manipulation. As both μ and σ are unknown, we evaluate this term by integrating out μ and σ where $f\left(\frac{\overline{y}_{t-1} - \mu_t}{\widetilde{s}}\right)$ has a Student t(t) distribution. Hence, we can estimate σ_d^2 by

$$\begin{aligned} \widehat{\sigma}_d^2 &= \frac{\nu^2 + t}{\nu^2 + t + 1}\left(s^2 + \delta^2 p_t\left(1 - p_t\right) + 2\delta p_t \widetilde{s} f\left(\frac{\overline{y}_{t-1} - \mu_t}{\widetilde{s}}\right)\right) \\ &= \frac{\nu^2 + t}{\nu^2 + t + 1}\left(\widetilde{s}^2 + 2\delta p_t \widetilde{s} f\left(\frac{\overline{y}_{t-1} - \mu_t}{\widetilde{s}}\right)\right) \end{aligned}$$

conditional on the data $\{y_t\}$.

Experiment

Repeat the experiment above except now we account for variance uncertainty.

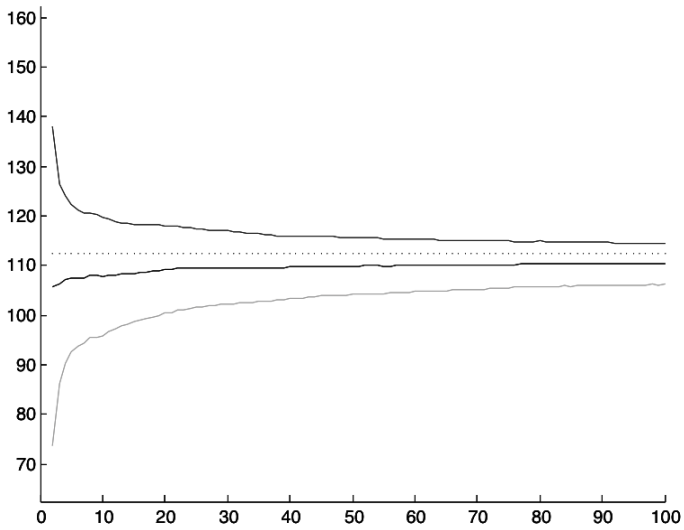

Figure 13.9: Selective manipulation σ_d unknown

Results

For 100 simulated samples of $t = 100$ observations, we generate a plot, figure 13.9, of the mean and average 95% interval estimates to summarize the data.As expected, the mean estimates converge toward the baseline (dashed line) as t increases and the interval estimates narrow around the baseline but not as rapidly as the known variance setting.

13.13 Jaynes' A_p distribution

Our story is nearly complete. However, consistent reasoning regarding propositions involves another, as yet unaddressed, element. For clarity, consider binary propositions. We might believe the propositions are equally likely but we also may be very confident of these probabilities, somewhat confident, or not confident at all. Jaynes [2003, ch. 18] compares propositions regarding heads or tails from a coin flip with life ever existing on Mars. He suggests that the former is very stable in light of additional evidence while the latter is very instable when faced with new evidence. Jaynes proposes a self-confessed odd proposition or distrib-

ution (depending on context) denoted A_p to tidily handle consistent reasoning.[47] The result is tidy in that evaluation of new evidence based on background knowledge (including A_p) follows from standard rules of probability theory — Bayes' theorem.

This new proposition is defined by

$$\Pr(A \mid A_p, E, \Im) \equiv p$$

where A is the proposition of interest, E is any additional evidence, $\Im$ mathematically relevant background knowledge, and A_p is something like regardless of anything else the probability of A is p. The propositions are mutually exclusive and exhaustive. As this is surely an odd proposition or distribution over a probability, let the distribution for A_p be denoted (A_p). High instability or complete ignorance leads to

$$(A_p \mid \Im) = 1 \quad 0 \leq p \leq 1$$

Bayes' theorem leads to

$$\begin{aligned}(A_p \mid E, \Im) &= (A_p \mid \Im)\frac{\Pr(E \mid A_p, \Im)\Pr(\Im)}{\Pr(E \mid \Im)\Pr(\Im)} \\ &= (A_p \mid \Im)\frac{\Pr(E \mid A_p, \Im)}{\Pr(E \mid \Im)}\end{aligned}$$

Given complete ignorance, this simplifies as

$$\begin{aligned}(A_p \mid E, \Im) &= (1)\frac{\Pr(E \mid A_p, \Im)}{\Pr(E \mid \Im)} \\ &= \frac{\Pr(E \mid A_p, \Im)}{\Pr(E \mid \Im)}\end{aligned}$$

Also, integrating out A_p we have

$$\Pr(A \mid E, \Im) = \int_0^1 (A, A_p \mid E, \Im)\, dp$$

expanding the integrand gives

$$\Pr(A \mid E, \Im) = \int_0^1 \Pr(A \mid A_p, E, \Im)\,(A_p \mid E, \Im)\, dp$$

from the definition of A_p, the first factor is simply p, leading to

$$\Pr(A \mid E, \Im) = \int_0^1 p \times (A_p \mid E, \Im)\, dp$$

Hence, the probability assigned to the proposition A is just the first moment or expected value of the distribution for A_p conditional on the new evidence. The key feature involves accounting for our uncertainty via the joint behavior of the prior and the likelihood.

---

[47] Jaynes' A_p distribution is akin to over-dispersed models. That is, hierarchical generalized linear models that allow dispersion beyond the assigned sampling distribution.

13.13.1 Football game puzzle revisited

Reconsider the football game puzzle posed by Walley [1991, pp. 270-271]. Recall the puzzle involves a football match-up between two football rivals which produces either a win (W), a loss (L), or a draw (D) for the home team. Suppose we know the home team wins *more than* half the time and we gain evidence the game doesn't end in a draw. Utilizing Jaynes' A_p distribution, the posterior distribution differs from the earlier case where the prior probability of a win is one-half, $\Pr(W, L, D) = \left(\frac{3}{4}, \frac{1}{4}, 0\right)$. The reasoning for this is as follows. Let A be the proposition the home team wins (the argument applies analogously to a loss) and we know only the probability is at least one-half, then

$$(A_p \mid \Im_1) = 2 \quad \tfrac{1}{2} \leq p \leq 1$$

and

$$(A_p \mid E, \Im_1) = (2)\, \frac{\Pr(E \mid A_p, \Im)}{\Pr(E \mid \Im)}$$

Since, $\Pr(E = \text{not } D \mid \Im_1) = \Pr(E = \text{not } D \mid A_p, \Im_1) = \frac{3}{4}$ if draws are permitted, or $\Pr(E = \text{not } D \mid \Im_2) = \Pr(E = \text{not } D \mid A_p, \Im_2) = 1$ if draws are not permitted by the game structure.

$$\begin{aligned} (A_p \mid E, \Im_1) &= (2)\, \frac{\frac{3}{4}}{\frac{3}{4}} = 2 \\ (A_p \mid E, \Im_2) &= (2)\, \frac{1}{1} = 2 \end{aligned}$$

Hence,

$$\begin{aligned} \Pr(A = W \mid E, \Im_j) &= \int_{\frac{1}{2}}^{1} p \cdot (A_p \mid E, \Im_j)\, dp \\ &= \int_{\frac{1}{2}}^{1} (2p)\, dp = \frac{3}{4} \end{aligned}$$

Here the puzzle is resolved by careful interpretation of prior uncertainty combined with consistent reasoning enforced by Jaynes' A_p distribution.[48] Prior instability forces us to reassess the evaluation of new evidence; consistent evaluation of the evidence is the key. Some alternative characterizations of our confidence in the prior probability the home team wins are illustrated next.

How might we reconcile Jaynes' A_p distribution and Walley's $\left\{\frac{2}{3}, \frac{1}{3}, 0\right\}$ or $\left\{\frac{1}{2}, \frac{1}{2}, 0\right\}$ probability conclusion. The former follows from background knowledge that the home team wins more than half the time with one-half most likely

[48] For a home team loss, we have

$$\Pr(A = L \mid E, \Im) = \int_0^{\frac{1}{2}} 2p dp = \frac{1}{4}$$

and monotonically declining toward one. A_p in this case is triangular $8 - 8p$ for $\frac{1}{2} \leq p \leq 1$. The latter case is supported by background knowledge that the home team wins about half the time but no other information regarding confidence in this claim. Then, A_p is uniform for $0 \leq p \leq 1$.

13.14 Concluding remarks

Now that we're "fully" armed, it's time to re-explore the accounting settings in this and previous chapters as well as other settings, collect data, and get on with the serious business of evaluating accounting choice. But this monograph must end somewhere, so we hope the reader will find continuation of this project a worthy task. We anxiously await the blossoming of an evidentiary archive and new insights.

13.15 Additional reading

There is a substantial and growing literature on maximum entropy priors. Jaynes [2003] is an excellent starting place. Cover and Thomas [1991, ch. 12] expand the maximum entropy principle via minimization of relative entropy in the form of a conditional limit theorem. Also, Cover and Thomas [1991, ch. 11] discuss maximum entropy distributions for time series data including Burg's theorem (Cover and Thomas [1991], pp. 274-5) stating the Gaussian distribution is the maximum entropy error distribution given autocovariances. Walley [1991] critiques the precise probability requirement of Bayesian analysis, the potential for improper ignorance priors, and the maximum entropy principle while arguing in favor of an upper and lower probability approach to consistent reasoning (see Jaynes' [2003] comment in the bibliography).

Financial statement inferences are extended to bounding transactions amounts and financial ratios in Arya, Fellingham, Mittendorf, and Schroeder [2004]. Earnings management implications for performance evaluation are discussed in path breaking papers by Arya, Glover, and Sunder [1998] and Demski [1998]. Arya et al discuss earnings management as a potential substitute for (lack of) commitment in conveying information about the manager's input. Demski discusses accruals smoothing as a potential means of conveying valuable information about the manager's talent and input. Demski, Fellingham, Lin, and Schroeder [2008] discuss the corrosive effects on organizations of excessive reliance on individual performance measures.

13.16 Appendix

This appendix supplies proofs to the propositions and corollaries for the smooth accruals discussion.

Proposition 13.1. *Let* $m_t = g\ m_{t-1} + \varepsilon_t$, $\Sigma = D$, *and* $\nu = \frac{\sigma_e}{\sigma_\epsilon}$. *Then,* $accruals_{t-1}$ *and* cf_t *are, collectively, sufficient statistics for the mean of cash flows* m_t *based on the history of cash flows and* $g^{t-1}accruals_t$ *is an efficient statistic for* m_t

$$\begin{aligned}[\widehat{m}_t|cf_1,...,cf_t] &= g^{t-1}accruals_t \\ &= \frac{1}{den_t}\left\{\frac{num_t}{g^2}(cf_t - a_t^*) + g^{t-1}\nu^2 den_{t-1}accruals_{t-1}\right\}\end{aligned}$$

where $accruals_0 = m_0$, *and* $\left[\begin{array}{c} den_t \\ num_t \end{array}\right] = B^t \left[\begin{array}{c} den_0 \\ num_0 \end{array}\right] = S\Lambda^t S^{-1}\left[\begin{array}{c} 1 \\ 0 \end{array}\right]$. *The variance of accruals is equal to the variance of the estimate of the mean of cash flows multiplied by* $g^{2(t-1)}$*; the variance of the estimate of the mean of cash flows equals the coefficient on current cash flow multiplied by* σ_e^2, $Var\left[\widehat{m}_t\right] = \frac{num_t}{den_t g^2}\sigma_e^2$.

Proof. Outline of the proof:

1. Since the data are multivariate normally distributed, *BLU* estimation is efficient (achieves the Cramer-Rao lower bound amongst consistent estimators; see Greene [1997], p. 300-302).
2. *BLU* estimation is written as a recursive least squares exercise (see Strang [1986], p. 146-148).
3. The proof is completed by induction. That is, the difference equation solution is shown, by induction, to be equivalent to the recursive least squares estimator. A key step is showing that the information matrix $\Im$ and its inverse can be derived in recursive fashion via LDL^T decomposition (i.e., $D^{-1}L^{-1}\Im = L^T$).

Recursive least squares. Let $H_1 = \left[\begin{array}{c} -\nu \\ 1 \end{array}\right]$ (a 2 by 1 matrix), $H_2 = \left[\begin{array}{cc} g\nu & -\nu \\ 0 & 1 \end{array}\right]$ (a 2 by 2 matrix), $H_t = \left[\begin{array}{ccccc} 0 & \cdots & 0 & g\nu & -\nu \\ 0 & \cdots & 0 & 0 & 1 \end{array}\right]$ (a 2 by t matrix with $t-2$ leading columns of zeroes), $z_1 = \left[\begin{array}{c} -g\nu m_0 \\ cf_1 - a_1^* \end{array}\right]$, $z_2 = \left[\begin{array}{c} 0 \\ cf_2 - a_2^* \end{array}\right]$, and $z_t = \left[\begin{array}{c} 0 \\ cf_t - a_t^* \end{array}\right]$. The information matrix for a t-period cash flow history is

$$\begin{aligned}\Im_t &= \Im_{t-1}^a + H_t^T H_t \\ &= \left[\begin{array}{ccccc} 1+\nu^2+g^2\nu^2 & -g\nu^2 & 0 & \cdots & 0 \\ -g\nu^2 & 1+\nu^2+g^2\nu^2 & -g\nu^2 & \ddots & \vdots \\ 0 & -g\nu^2 & \ddots & -g\nu^2 & 0 \\ \vdots & \ddots & -g\nu^2 & 1+\nu^2+g^2\nu^2 & -g\nu^2 \\ 0 & \cdots & 0 & -g\nu^2 & 1+\nu^2 \end{array}\right]\end{aligned}$$

a symmetric tri-diagonal matrix, where $\Im_{t-1}^a$ is $\Im_{t-1}$ augmented with a row and column of zeroes to conform with $\Im_t$. For instance, $\Im_1 = \left[1+\nu^2\right]$ and $\Im_1^a = \begin{bmatrix} 1+\nu^2 & 0 \\ 0 & 0 \end{bmatrix}$. The estimate of the mean of cash flows is derived recursively as

$$b_t = b_{t-1}^a + k_t\left(z_t - H_t b_{t-1}^a\right)$$

for $t > 1$ where $k_t = \Im_t^{-1} H_t^T$, the gain matrix, and b_{t-1}^a is b_{t-1} augmented with a zero to conform with b_t. The best linear unbiased estimate of the current mean is the last element in the vector b_t and its variance is the last row-column element of $\Im_t^{-1}$ multiplied by σ_e^2.

Difference equations. The difference equations are

$$\begin{bmatrix} den_t \\ num_t \end{bmatrix} = \begin{bmatrix} 1+\nu^2 & \nu^2 \\ g^2 & g^2\nu^2 \end{bmatrix} \begin{bmatrix} den_{t-1} \\ num_{t-1} \end{bmatrix}$$

with $\begin{bmatrix} den_0 \\ num_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. The difference equations estimator for the current mean of cash flows and its variance are

$$\begin{aligned} \widehat{m}_t &= \frac{1}{den_t}\left(\frac{num_t}{g^2}\left(cf_t - a_t^*\right) + g\nu^2 den_{t-1}\widehat{m}_{t-1}\right) \\ &= g^{t-1}\, accruals_t \\ &= \frac{1}{den_t}\left(\frac{num_t}{g^2}\left(cf_t - a_t^*\right) + g^{t-1}\nu^2 den_{t-1} accruals_{t-1}\right) \end{aligned}$$

where $accruals_0 = m_0$, and

$$Var\left[\widehat{m}_t\right] = g^{2(t-1)} Var\left[accruals_t\right] = \sigma_e^2 \frac{num_t}{g^2 den_t}.$$

Induction steps. Assume

$$\begin{aligned} \widehat{m}_t &= \frac{1}{den_t}\left(\frac{num_t}{g^2}\left(cf_t - a_t^*\right) + g\nu^2 den_{t-1}\widehat{m}_{t-1}\right) \\ &= g^{t-1}\, accruals_t \\ &= \frac{1}{den_t}\left(\frac{num_t}{g^2}\left(cf_t - a_t^*\right) + g^{t-1}\nu^2 den_{t-1} accruals_{t-1}\right) \\ &= \left[b_{t-1}^a + k_t\left(z_t - H_t b_{t-1}^a\right)\right][t] \end{aligned}$$

and

$$Var\left[\widehat{m}_t\right] = g^{2(t-1)} Var\left[accruals_t\right] = Var\left[b_t\right][t,t]$$

where $[t]$ ($[t,t]$) refers to element t (t,t) in the vector (matrix). The above is clearly true for the base case, $t = 1$ and $t = 2$. Now, show

$$\begin{aligned} \widehat{m}_{t+1} &= \frac{1}{den_{t+1}}\left(\frac{num_{t+1}}{g^2}\left(cf_{t+1} - a_{t+1}^*\right) + g^t\nu^2 den_t accruals_t\right) \\ &= \left[b_t^a + k_{t+1}\left(z_{t+1} - H_{t+1} b_t^a\right)\right][t+1]\,. \end{aligned}$$

Recall $z_{t+1} = \begin{bmatrix} 0 \\ c_{t+1} - a^*_{t+1} \end{bmatrix}$ and $H_{t+1} = \begin{bmatrix} 0 & \cdots & 0 & g\nu & -\nu \\ 0 & \cdots & 0 & 0 & 1 \end{bmatrix}$. From LDL^T decomposition of $\Im_{t+1}$ (recall $L^T = D^{-1}L^{-1}\Im$ where L^{-1} is simply products of matrices reflecting successive row eliminations - no row exchanges are involved due to the tri-diagonal structure and D^{-1} is the reciprocal of the diagonal elements remaining following eliminations) the last row of $\Im^{-1}_{t+1}$ is

$$\begin{bmatrix} \frac{g^{t-1}\nu^{2(t-1)}num_1}{g^2den_{t+1}} & \cdots & \frac{g^2\nu^4 num_{t-1}}{g^2den_{t+1}} & \frac{g\nu^2 num_t}{g^2den_{t+1}} & \frac{num_{t+1}}{g^2den_{t+1}} \end{bmatrix}.$$

This immediately identifies the variance associated with the estimator as the last term in $\Im^{-1}_{t+1}$ multiplied by the variance of cash flows, $\frac{num_{t+1}}{g^2den_{t+1}}\sigma^2_e$. Hence, the difference equation and the recursive least squares variance estimators are equivalent.

Since $H^T_{t+1}z_{t+1} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ cf_{t+1} - a^*_{t+1} \end{bmatrix}$, the lead term on the RHS of the $[t+1]$ mean estimator is $\frac{num_{t+1}}{g^2den_{t+1}}\left(cf_{t+1} - a^*_{t+1}\right)$ which is identical to the lead term on the left hand side (LHS). Similarly, the second term on the RHS (recall the focus is on element t, the last element of b^a_t is 0) is

$$\begin{aligned}
& \left[\left(I - k_{t+1}H_{t+1}\right)b^a_t\right][t+1] \\
= & \left[\left(I - \Im^{-1}_{t+1}\begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & \vdots & \ddots & \vdots & \vdots \\ \vdots & 0 & \ddots & 0 & 0 \\ 0 & \cdots & 0 & g^2\nu^2 & -g\nu^2 \\ 0 & \cdots & 0 & -g\nu^2 & 1+\nu^2 \end{bmatrix}\right) b^a_t\right][t+1] \\
= & \left(\frac{-g^3\nu^4 num_t}{g^2den_{t+1}} + \frac{g\nu^2 num_{t+1}}{g^2den_{t+1}}\right)\widehat{m}_t \\
= & \left(\frac{-g^3\nu^4 num_t + g\nu^2 num_{t+1}}{g^2den_{t+1}}\right) g^{t-1}\, accruals_t.
\end{aligned}$$

The last couple of steps involve substitution of $\widehat{m}_t$ for $b^a_t[t+1]$ followed by $g^{t-1}accruals_t$ for $\widehat{m}_t$ on the right hand side (RHS) The difference equation relation, $num_{t+1} = g^2den_t + g^2\nu^2 num_t$, implies

$$\begin{aligned}
\frac{-g^3\nu^4 num_t + g\nu^2 num_{t+1}}{g^2den_{t+1}}\widehat{m}_t &= \frac{1}{den_{t+1}} g\nu^2 den_t\, \widehat{m}_t \\
&= \frac{1}{den_{t+1}} g^t\nu^2 den_t\, accruals_t
\end{aligned}$$

the second term on the LHS. This completes the induction steps. ■

Corollary 13.2. *As t becomes large, the weight on current cash flows for the efficient estimator of the mean of cash flows approaches*

$$\frac{2}{1+\left(1-g^2\right)\nu^2+\sqrt{\left(1+\left(1+g^2\right)\nu^2\right)^2-4g^2\nu^4}}$$

and the variance of the estimate approaches

$$\frac{2}{1+\left(1-g^2\right)\nu^2+\sqrt{\left(1+\left(1+g^2\right)\nu^2\right)^2-4g^2\nu^4}}\sigma_e^2.$$

Proof. The difference equations

$$\begin{aligned}\begin{bmatrix} den_t \\ num_t \end{bmatrix} &= S\Lambda^t S^{-1}\begin{bmatrix} den_0 \\ num_0 \end{bmatrix} \\ &= S\Lambda^t S^{-1}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = S\Lambda^t c\end{aligned}$$

imply

$$c = S^{-1}\begin{bmatrix} den_0 \\ num_0 \end{bmatrix} = \begin{bmatrix} \frac{-g^2}{1+(1+g^2)\nu^2+\sqrt{(1+(1+g^2)\nu^2)^2-4g^2\nu^4}} \\ \frac{g^2}{1+(1+g^2)\nu^2+\sqrt{(1+(1+g^2)\nu^2)^2-4g^2\nu^4}} \end{bmatrix}$$

Thus,

$$\begin{bmatrix} den_t \\ num_t \end{bmatrix} = S\begin{bmatrix} \lambda_1^t & 0 \\ 0 & \lambda_2^t \end{bmatrix} c$$

$$\begin{aligned}= &\frac{1}{\sqrt{\left(1+\left(1+g^2\right)\nu^2\right)^2-4g^2\nu^4}} \\ &\times\begin{bmatrix} \frac{1}{2}\left\{\lambda_2^t\left(1+\left(1-g^2\right)\nu^2+\sqrt{\left(1+\left(1+g^2\right)\nu^2\right)^2-4g^2\nu^4}\right)\right. \\ \left.-\lambda_1^t\left(1+\left(1-g^2\right)\nu^2-\sqrt{\left(1+\left(1+g^2\right)\nu^2\right)^2-4g^2\nu^4}\right)\right\} \\ g^2\left(\lambda_2^t-\lambda_1^t\right)\end{bmatrix}\end{aligned}$$

Since λ_2 is larger than λ_1, λ_1^t contributes negligibly to $\begin{bmatrix} den_t \\ num_t \end{bmatrix}$ for arbitrarily large t. Hence,

$$\lim_{t\to\infty}\frac{num_t}{g^2 den_t} = \frac{2}{1+\left(1-g^2\right)\nu^2+\sqrt{\left(1+\left(1+g^2\right)\nu^2\right)^2-4g^2\nu^4}}.$$

From proposition 13.1, the variance of the estimator for expected cash flow is $\frac{num_t}{g^2 den_t}\sigma_e^2$. Since

$$\lim_{t\to\infty}\frac{num_t}{g^2 den_t} = \frac{2}{1+(1-g^2)\,\nu^2+\sqrt{\left(1+(1+g^2)\,\nu^2\right)^2-4g^2\nu^4}}.$$

the asymptotic variance is

$$\frac{2}{1+(1-g^2)\,\nu^2+\sqrt{\left(1+(1+g^2)\,\nu^2\right)^2-4g^2\nu^4}}\sigma_e^2.$$

This completes the asymptotic case. ■

Proposition 13.2. *Let* $m_t = g\,m_{t-1}+\varepsilon_t$, $\Sigma = D$, $\nu = \frac{\sigma_e}{\sigma_\epsilon}$, *and* $\phi = \frac{\sigma_e}{\sigma_\mu}$. *Then,* $accruals_{t-1}$, cf_t, *and* y_t, *collectively, are sufficient statistics for evaluating the agent with incentive payments given by*

$$\begin{aligned}\gamma_t^T w_t &= \Delta\frac{1}{\nu^2 den_{t-1}+\phi^2 den_t}\\ &\quad\times\left[\phi^2 den_t y_t+\nu^2 den_{t-1}\left(cf_t-g^{t-1}accruals_{t-1}\right)\right]\end{aligned}$$

and variance of payments equal to

$$Var[\gamma_t^T w_t] = \Delta^2\frac{den_t}{\nu^2 den_{t-1}+\phi^2 den_t}\sigma_e^2$$

where $\Delta = \frac{c(a_H)-c(a_L)}{aH-aL}$, and $accruals_{t-1}$ *and* den_t *are as defined in proposition 13.1.*

Proof. Outline of the proof:

1. Show that the "best" linear contract is equivalent to the *BLU* estimator of the agent's current act rescaled by the agent's marginal cost of the act.
2. The *BLU* estimator is written as a recursive least squares exercise (see Strang [1986], p. 146-148).
3. The proof is completed by induction. That is, the difference equation solution is shown, by induction, to be equivalent to the recursive least squares estimator. Again, a key step involves showing that the information matrix $\Im_a$ and its inverse can be derived in recursive fashion via LDL^T decomposition (i.e., $D^{-1}L^{-1}\Im_a = L^T$).

"Best" linear contracts. The program associated with the optimal a_H-inducing *LEN* contract written in certainty equivalent form is

$$\underset{\delta,\gamma}{Min}\ \delta + E\left[\gamma^T w|a_H\right]$$

subject to

$$\delta + E\left[\gamma^T w|a_H\right] - \frac{r}{2}Var\left[\gamma^T w\right] - c\left(a_H\right) \geq RW \qquad \text{(IR)}$$

$$\delta + E\left[\gamma^T w|a_H\right] - \frac{r}{2}Var\left[\gamma^T w\right] - c\left(a_H\right) \geq \delta + E\left[\gamma^T w|a_L\right] - \frac{r}{2}Var\left[\gamma^T w\right] - c\left(a_L\right) \qquad \text{(IC)}$$

As demonstrated in Arya et al [2004], both *IR* and *IC* are binding and γ equals the *BLU* estimator of a based on the history w (the history of cash flows cf and other contractible information y) rescaled by the agent's marginal cost of the act $\Delta = \frac{c(a_H)-c(a_L)}{a_H-a_L}$. Since *IC* is binding,

$$\begin{aligned} & \delta + E\left[\gamma^T w|a_H\right] - \frac{r}{2}Var\left[\gamma^T w\right] - \left(\delta + E\left[\gamma^T w|a_L\right] - \frac{r}{2}Var\left[\gamma^T w\right]\right) \\ = \quad & c\left(a_H\right) - c\left(a_L\right) \end{aligned}$$

$$E\left[\gamma^T w|a_H\right] - E\left[\gamma^T w|a_L\right] = c\left(a_H\right) - c\left(a_L\right)$$

$$\gamma^T\left\{E\left[w|a_H\right] - E\left[w|a_L\right]\right\} = c\left(a_H\right) - c\left(a_L\right)$$

$$\left(a_H - a_L\right)\gamma^T\hbar = c\left(a_H\right) - c\left(a_L\right)$$

where

$$w = \begin{bmatrix} cf_1 - m_0 - a_1^* \\ cf_2 - m_0 - a_2^* \\ \vdots \\ cf_t - m_0 \\ y_t \end{bmatrix}$$

and $\hbar$ is a vector of zeroes except the last two elements are equal to one, and

$$\gamma^T\hbar = \frac{c\left(a_H\right) - c\left(a_L\right)}{a_H - a_L}.$$

Notice, the sum of the last two elements of γ equals one, $\gamma^T\hbar = 1$, is simply the unbiasedness condition associated with the variance minimizing estimator of a based on design matrix H_a. Hence, $\gamma^T w$ equals the *BLU* estimator of a rescaled by Δ, $\gamma_t^T w_t = \Delta\widehat{a}_t$. As δ is a free variable, *IR* can always be exactly satisfied by setting

$$\delta = RW - \left\{E\left[\gamma^T w|a_H\right] - \frac{r}{2}Var\left[\gamma^T w\right] - c\left(a_H\right)\right\}.$$

Recursive least squares. H_t remains as defined in the proof of proposition 13.1. Let $H_{a1} = \begin{bmatrix} -\nu & 0 \\ 1 & 1 \\ 0 & \phi \end{bmatrix}$ (a 3 by 2 matrix), $H_{a2} = \begin{bmatrix} g\nu & -\nu & 0 \\ 0 & 1 & 1 \\ 0 & 0 & \phi \end{bmatrix}$ (a 3 by 3

matrix), $H_{at} = \begin{bmatrix} 0 & \cdots & 0 & g\nu & -\nu & 0 \\ 0 & \cdots & 0 & 0 & 1 & 1 \\ 0 & \cdots & 0 & 0 & 0 & \phi \end{bmatrix}$ (a 3 by $t+1$ matrix with leading zeroes), $\widetilde{w}_1 = \begin{bmatrix} -g\nu m_0 \\ cf_1 \\ y_1 \end{bmatrix}$, $\widetilde{w}_2 = \begin{bmatrix} 0 \\ cf_2 \\ y_2 \end{bmatrix}$, and $\widetilde{w}_t = \begin{bmatrix} 0 \\ cf_t \\ y_t \end{bmatrix}$. The information matrix for a t-period cash flow and other monitoring information history is

$$\Im_{at} = \Im_{t-1}^{aa} + H_{at}^T H_{at} =$$

$$\begin{bmatrix} 1+\nu^2+g^2\nu^2 & -g\nu^2 & 0 & 0 & \cdots & 0 \\ -g\nu^2 & 1+\nu^2+g^2\nu^2 & -g\nu^2 & \ddots & \cdots & 0 \\ 0 & -g\nu^2 & \ddots & \ddots & 0 & \vdots \\ 0 & \ddots & \ddots & 1+\nu^2+g^2\nu^2 & -g\nu^2 & 0 \\ \vdots & \cdots & 0 & -g\nu^2 & 1+\nu^2 & 1 \\ 0 & 0 & \cdots & 0 & 1 & 1+\phi^2 \end{bmatrix}$$

a symmetric tri-diagonal matrix where $\Im_{t-1}^{aa}$ is $\Im_{t-1}^{a}$ (the augmented information matrix employed to estimate the cash flow mean in proposition 13.1) augmented with an additional row and column of zeroes (i.e., the information matrix from proposition 13.1, $\Im_{t-1}$, is augmented with two columns of zeroes on the right and two rows of zeroes on the bottom). The recursive least squares estimator is

$$b_{at} = \left[b_{t-1}^{aa} + k_{at} \left(\widetilde{w}_t - H_{at} b_{t-1}^{aa} \right) \right]$$

for $t > 1$ where b_{t-1}^{aa} is b_{t-1} (the accruals estimator of m_{t-1} from proposition 13.1) augmented with two zeroes and $k_{at} = \Im_{at}^{-1} H_{at}^T$. The best linear unbiased estimate of the current act is the last element in the vector b_{at} and its variance is the last row-column element of $\Im_{at}^{-1}$ multiplied by σ_e^2. Notice, recursive least squares applied to the performance evaluation exercise utilizes the information matrix $\Im_{t-1}^{aa}$ (the information matrix employed in proposition 13.1 augmented with two trailing rows-columns of zeroes) and estimator b_{t-1}^{aa} (the accruals estimator of m_{t-1} from proposition 13.1 augmented with the two trailing zeroes). This accounts for the restriction on the parameters due to past actions already having been motivated in the past (i.e., past acts are at their equilibrium level a^*). Only the current portion of the design matrix H_{at} and the current observations w_t (in place of z_t) differ from the setup for accruals (in proposition 13.1).

Difference equations. The difference equations are

$$\begin{bmatrix} den_t \\ num_t \end{bmatrix} = \begin{bmatrix} 1+\nu^2 & \nu^2 \\ g^2 & g^2\nu^2 \end{bmatrix} \begin{bmatrix} den_{t-1} \\ num_{t-1} \end{bmatrix}$$

with $\begin{bmatrix} den_0 \\ num_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. The difference equations estimator for the linear incentive payments $\gamma^T w$ is

$$\begin{aligned}
\gamma_t^T w_t &= \Delta \frac{1}{\nu^2 den_{t-1} + \phi^2 den_t} \left[\phi^2 den_t y_t + \nu^2 den_{t-1} \left(cf_t - g\, \widehat{m}_{t-1}\right)\right] \\
&= \Delta \frac{1}{\nu^2 den_{t-1} + \phi^2 den_t} \\
&\quad \times \left[\phi^2 den_t y_t + \nu^2 den_{t-1} \left(cf_t - g^{t-1}\, accruals_{t-1}\right)\right]
\end{aligned}$$

and the variance of payments is

$$Var\left[\gamma^T w\right] = \Delta^2 \frac{den_t}{\nu^2 den_{t-1} + \phi^2 den_t} \sigma_e^2.$$

Induction steps. Assume

$$\begin{aligned}
\gamma_t^T w_t &= \Delta \frac{1}{\nu^2 den_{t-1} + \phi^2 den_t} \left[\phi^2 den_t y_t + \nu^2 den_{t-1} \left(cf_t - g\, \widehat{m}_{t-1}\right)\right] \\
&= \Delta \frac{1}{\nu^2 den_{t-1} + \phi^2 den_t} \\
&\quad \times \left[\phi^2 den_t y_t + \nu^2 den_{t-1} \left(cf_t - g^{t-1}\, accruals_{t-1}\right)\right] \\
&= \Delta \left[b_{t-1}^a + k_{at} \left(w_t - H_{at} b_{t-1}^a\right)\right][t+1]
\end{aligned}$$

and

$$Var\left[\gamma_t^T w_t\right] = \Delta^2 Var\left[\widehat{a}_t\right][t+1, t+1]$$

where $[t+1]$ ($[t+1, t+1]$) refers to element $t+1$ $(t+1, t+1)$ in the vector (matrix). The above is clearly true for the base case, $t = 1$ and $t = 2$. Now, show

$$\begin{aligned}
&\Delta \frac{1}{\nu^2 den_t + \phi^2 den_{t+1}} \left[\phi^2 den_{t+1} y_{t+1} + \nu^2 den_t \left(cf_{t+1} - g\, \widehat{m}_t\right)\right] \\
= \;& \Delta \frac{1}{\nu^2 den_t + \phi^2 den_{t+1}} \left[\phi^2 den_{t+1} y_{t+1} + \nu^2 den_t \left(cf_{t+1} - g^t\, accruals_t\right)\right] \\
= \;& \Delta \left[b_t^a + k_{at+1} \left(\widetilde{w}_{t+1} - H_{at+1} b_t^a\right)\right][t+2].
\end{aligned}$$

Recall $\widetilde{w}_{t+1} = \begin{bmatrix} 0 \\ cf_{t+1} \\ \phi y_{t+1} \end{bmatrix}$ and $H_{at+1} = \begin{bmatrix} 0 & \cdots & 0 & g\nu & \nu & 0 \\ 0 & \cdots & 0 & 0 & 1 & 1 \\ 0 & \cdots & 0 & 0 & 0 & \phi \end{bmatrix}$. From LDL^T decomposition of $\Im_{at+1}$ (recall $L^T = D^{-1} L^{-1} \Im_a$ where L^{-1} is simply products of matrices reflecting successive row eliminations - no row exchanges are involved due to the tri-diagonal structure and D^{-1} is the reciprocal of the

remaining elements remaining after eliminations) the last row of $\Im_{at+1}^{-1}$ is

$$\frac{1}{\nu^2 den_t + \phi^2 den_{t+1}} \begin{bmatrix} -g^{t-1}\nu^{2(t-1)} den_1 \\ \vdots \\ -g\nu^2 \left(den_{t-1} + \nu^2 num_{t-1} \right) \\ -\left(den_t + \nu^2 num_t \right) \\ den_{t+1} \end{bmatrix}^T .^{49}$$

This immediately identifies the variance associated with the estimator as the last term in $\Im_{at+1}^{-1}$ multiplied by the product of the agent's marginal cost of the act squared and the variance of cash flows, $\Delta^2 \frac{den_{t+1}}{\nu^2 den_t + \phi^2 den_{t+1}} \sigma_e^2$. Hence, the difference equation and the recursive least squares variance of payments estimators are equivalent.

Since $H_{at+1}^T \widetilde{w}_{t+1} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ cf_{t+1} \\ cf_{t+1} + y_{t+1} \end{bmatrix}$ and the difference equation implies $den_{t+1} = \left(1 + \nu^2\right) den_t + \nu^2 num_t$, the lead term on the RHS is

$$\begin{aligned} & \frac{den_{t+1}}{\nu^2 den_t + \phi^2 den_{t+1}} \left(y_{t+1} + cf_{t+1} \right) - \frac{den_t + \nu^2 num_t}{\nu^2 den_t + \phi^2 den_{t+1}} cf_{t+1} \\ = & \frac{den_{t+1}}{\nu^2 den_t + \phi^2 den_{t+1}} y_{t+1} - \frac{\nu^2 den_t}{\nu^2 den_t + \phi^2 den_{t+1}} cf_{t+1} \end{aligned}$$

which equals the initial expression on the LHS of the $[t+2]$ incentive payments. Similarly, the $\widehat{m}_t = g^{t-1}\, accruals_t$ term on the RHS (recall the focus is on element $t+2$) is

$$\begin{aligned} & \left[\left(I - k_{at+1} H_{at+1} \right) b_t^a \right] [t+2] \\ = & \left[\left(I - \Im_{at+1}^{-1} \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & g^2\nu^2 & -g\nu^2 & 0 \\ 0 & \cdots & 0 & -g\nu^2 & 1+\nu^2 & 1 \\ 0 & \cdots & 0 & 0 & 1 & 1+\phi^2 \end{bmatrix} \right) b_t^a \right] [t+2] \\ = & -\frac{g\nu^2 den_t}{\nu^2 den_t + \phi^2 den_{t+1}} \widehat{m}_t \\ = & -\frac{g^t \nu^2 den_t}{\nu^2 den_t + \phi^2 den_{t+1}} accruals_t. \end{aligned}$$

[49] Transposed due to space limitations.

Combining terms and simplifying produces the result

$$\begin{aligned} & \frac{1}{\nu^2 den_t + \phi^2 den_{t+1}} \left[\phi^2 den_{t+1} y_{t+1} + \nu^2 den_t \left(cf_{t+1} - g\, \widehat{m}_t\right)\right] \\ = \quad & \frac{1}{\nu^2 den_t + \phi^2 den_{t+1}} \left[\phi^2 den_{t+1} y_{t+1} + \nu^2 den_t \left(cf_{t+1} - g^t\, accruals_t\right)\right]. \end{aligned}$$

Finally, recall the estimator $\widehat{a}_t$ (the last element of b_{at}) rescaled by the agent's marginal cost of the act identifies the "best" linear incentive payments

$$\begin{aligned} \gamma_t^T w_t &= \Delta \widehat{a}_t \\ &= \Delta \frac{1}{\nu^2 den_{t-1} + \phi^2 den_t} \left[\phi^2 den_t y_t + \nu^2 den_{t-1} \left(cf_t - g\, \widehat{m}_{t-1}\right)\right] \\ &= \Delta \frac{1}{\nu^2 den_{t-1} + \phi^2 den_t} \\ &\quad \times \left[\phi^2 den_t y_t + \nu^2 den_{t-1} \left(cf_t - g^{t-1}\, accruals_{t-1}\right)\right]. \end{aligned}$$

This completes the induction steps. ■

Corollary 13.4. *For the benchmark case $\Sigma = \sigma_e^2 I$ ($\nu = \phi = 1$) and $g = 1$, accruals at time t are an efficient summary of past cash flow history for the cash flow mean if*

$$\begin{aligned} [\widehat{m}_t | cf_1, ..., cf_t] &= accruals_t \\ &= \frac{F_{2t}}{F_{2t+1}} \left(cf_t - a_t^*\right) + \frac{F_{2t-1}}{F_{2t+1}} accruals_{t-1} \end{aligned}$$

where $F_n = F_{n-1} + F_{n-2}$, $F_0 = 0$, $F_1 = 1$ (the Fibonacci series), and the sequence is initialized with $accruals_0 = m_0$ (common knowledge mean beliefs). Then,variance of accruals equals $Var\left[\widehat{m}_t\right] = \frac{F_{2t}}{F_{2t+1}} \sigma_e^2$.

Proof. Replace $g = \nu = 1$ in proposition 13.1. Hence,

$$\begin{bmatrix} den_t \\ num_t \end{bmatrix} = B \begin{bmatrix} den_{t-1} \\ num_{t-1} \end{bmatrix}$$

reduces to

$$\begin{bmatrix} den_t \\ num_t \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} den_{t-1} \\ num_{t-1} \end{bmatrix}.$$

Since

$$\begin{bmatrix} F_{n+1} \\ F_n \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix}$$

and

$$\begin{bmatrix} F_{n+2} \\ F_{n+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix},$$

$$den_t = F_{2t+1}, num_t = F_{2t}, den_{t-1} = F_{2t-1}, \text{and } num_{t-1} = F_{2t-2}.$$

For $g = \nu = 1$, the above implies

$$\begin{aligned} \widehat{m}_t &= g^{t-1} accruals_t \\ &= \frac{1}{den_t}\left(\frac{num_t}{g^2}\left(cf_t - a_t^*\right) + g^{t-1}\nu^2 den_{t-1} accruals_{t-1}\right) \end{aligned}$$

reduces to

$$\frac{F_{2t}}{F_{2t+1}}\left(cf_t - a_t^*\right) + \frac{F_{2t-1}}{F_{2t+1}} accruals_{t-1}$$

and variance of accruals equals $\frac{F_{2t}}{F_{2t+1}}\sigma_e^2$. ■

Corollary 13.5 *For the benchmark case* $\Sigma = \sigma_e^2 I$ $(\nu = \phi = 1)$ *and* $g = 1$,$accruals_{t-1}$, cf_t, *and* y_t *are, collectively, sufficient statistics for evaluating the agent with incentive payments given by*

$$\gamma_t^T w_t = \Delta\left\{\frac{F_{2t+1}}{L_{2t}} y_t + \frac{F_{2t-1}}{L_{2t}}\left(cf_t - accruals_{t-1}\right)\right\}$$

and variance of payments equals $\Delta^2 \frac{F_{2t+1}}{L_{2t}}\sigma_e^2$ *where* $accruals_{t-1}$ *is as defined in corollary 13.4 and* $L_n = L_{n-1} + L_{n-2}$, $L_0 = 2$, *and* $L_1 = 1$ *(the Lucas series), and* $\Delta = \frac{c(a_H) - c(a_L)}{a_H - a_L}$.

Proof. Replace $g = \nu = \phi = 1$ in proposition 13.3. Hence,

$$\begin{bmatrix} den_t \\ num_t \end{bmatrix} = B \begin{bmatrix} den_{t-1} \\ num_{t-1} \end{bmatrix}$$

reduces to

$$\begin{bmatrix} den_t \\ num_t \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} den_{t-1} \\ num_{t-1} \end{bmatrix}.$$

Since

$$\begin{bmatrix} F_{n+1} \\ F_n \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix}$$

and

$$\begin{bmatrix} F_{n+2} \\ F_{n+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix}$$

$den_t = F_{2t+1}$, $num_t = F_{2t}$, $den_{t-1} = F_{2t-1}$, $num_{t-1} = F_{2t-2}$, and $L_t = F_{t+1} + F_{t-1}$. For $g = \nu = \phi = 1$, the above implies

$$\gamma_t^T w_t = \Delta \frac{1}{\nu^2 den_{t-1} \phi^2 den_t}\left[\phi^2 den_t y_t + \nu^2 den_{t-1}\left(cf_t - g^{t-1} accruals_{t-1}\right)\right]$$

reduces to

$$\Delta\left\{\frac{F_{2t-1}}{L_{2t}}\left(cf_t - accruals_{t-1}\right) + \frac{F_{2t+1}}{L_{2t}} y_t\right\}$$

and variance of payments equals $\Delta^2 \frac{F_{2t+1}}{L_{2t}}\sigma_e^2$. ■