Introduction to Bayes Linear Statistics

Jonathan Cumming, Ian Vernon

3rd September, 2010

Introduction

Introduction

Bayesian Inference

- Bayesian inference is a method of statistical inference in which we use data to update beliefs about uncertain quantities of interest.

Bayesian Inference

- Bayesian inference is a method of statistical inference in which we use data to update beliefs about uncertain quantities of interest.
- In this methodology, our uncertainty about any quantities of interest is quantified by probability distributions

Bayesian Inference

- Bayesian inference is a method of statistical inference in which we use data to update beliefs about uncertain quantities of interest.
- In this methodology, our uncertainty about any quantities of interest is quantified by probability distributions
- Our updated beliefs about the quantity of interest, θ, given the data, D, are then obtained via application of Bayes Theorem:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Bayesian Inference

- $p(\theta)$ represents our beliefs about θ before the data D become available - the prior for θ.

Bayesian Inference

- $p(\theta)$ represents our beliefs about θ before the data D become available - the prior for θ.
- $p(D|\theta)$ is the probability of the data given the uncertain quantity θ

Bayesian Inference

- $p(\theta)$ represents our beliefs about θ before the data D become available - the prior for θ.
- $p(D|\theta)$ is the probability of the data given the uncertain quantity θ
- $p(\theta|D)$ represents our beliefs about θ after the data D have been observed - the posterior for θ

Bayesian Inference

- $p(\theta)$ represents our beliefs about θ before the data D become available - the prior for θ.
- $p(D|\theta)$ is the probability of the data given the uncertain quantity θ
- $p(\theta|D)$ represents our beliefs about θ after the data D have been observed - the posterior for θ
- These first three elements are the heart of Bayesian inference, which can be remembered as

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Some interesting questions

- We need to specify a full joint probability distribution for all uncertain quantities
  - Specifying a probability distribution is equivalent to specifying an infinite set of moments, do we really believe all of those implicit specifications?

Some interesting questions

- We need to specify a full joint probability distribution for all uncertain quantities
  - Specifying a probability distribution is equivalent to specifying an infinite set of moments, do we really believe all of those implicit specifications?
- When the prior and likelihood have specific forms, then this posterior can be found analytically (conjugacy)
  - Are those distributions really REALLY conjugate, or are they just convenient?

Some interesting questions

- We need to specify a full joint probability distribution for all uncertain quantities
  - Specifying a probability distribution is equivalent to specifying an infinite set of moments, do we really believe all of those implicit specifications?
- When the prior and likelihood have specific forms, then this posterior can be found analytically (conjugacy)
  - Are those distributions really REALLY conjugate, or are they just convenient?
- In all other cases, we must rely on intensive computational methods to arrive a distribution for $p(\theta|D)$
  - If we don't completely believe our prior specification, what faith should we have in this posterior?

Difficulties with full Bayes

- Working within a fully Bayesian framework, we can encounter certain difficulties when considering multivariate analyses

Difficulties with full Bayes

- Working within a fully Bayesian framework, we can encounter certain difficulties when considering multivariate analyses
- Even in small problems, it can be extremely difficult and/or time-consuming to distil all the prior knowledge into a meaningful joint prior probability specification;

Difficulties with full Bayes

- Working within a fully Bayesian framework, we can encounter certain difficulties when considering multivariate analyses
- Even in small problems, it can be extremely difficult and/or time-consuming to distil all the prior knowledge into a meaningful joint prior probability specification;
- Even with a specification, the computations for learning from data become both difficult and computer intensive;

Difficulties with full Bayes

- Working within a fully Bayesian framework, we can encounter certain difficulties when considering multivariate analyses
- Even in small problems, it can be extremely difficult and/or time-consuming to distil all the prior knowledge into a meaningful joint prior probability specification;
- Even with a specification, the computations for learning from data become both difficult and computer intensive;
- In higher-dimensions the likelihood surface can be very complicated, making full Bayes calculations potentially highly non-robust.

Difficulties with full Bayes

- Working within a fully Bayesian framework, we can encounter certain difficulties when considering multivariate analyses
- Even in small problems, it can be extremely difficult and/or time-consuming to distil all the prior knowledge into a meaningful joint prior probability specification;
- Even with a specification, the computations for learning from data become both difficult and computer intensive;
- In higher-dimensions the likelihood surface can be very complicated, making full Bayes calculations potentially highly non-robust.
- Therefore if we are unable to make and analyse full prior probability specifications, we require methods based around simpler belief specifications

Expectation as Primitive

- Rather than work with probability as the fundamental quantity of uncertainty, we could use expectation
- de Finetti spent most of his life studying subjective conceptions of probability.
- He proposed the use of expectation as the primitive entity on which to base any analysis, as opposed to probability.

Expectation as Primitive

- Rather than work with probability as the fundamental quantity of uncertainty, we could use expectation
- de Finetti spent most of his life studying subjective conceptions of probability.
- He proposed the use of expectation as the primitive entity on which to base any analysis, as opposed to probability.
- Probabilities (where relevant) enter as derived quantities: they are the expectations of indicator functions.

Expectation as Primitive

- Rather than work with probability as the fundamental quantity of uncertainty, we could use expectation
- de Finetti spent most of his life studying subjective conceptions of probability.
- He proposed the use of expectation as the primitive entity on which to base any analysis, as opposed to probability.
- Probabilities (where relevant) enter as derived quantities: they are the expectations of indicator functions.
- Note this asymmetry: if probability is treated as the primitive quantity then one has to specify (in the continuous case) an infinite set of probabilities in order to derive a single expectation.

Expectation as Primitive

- Rather than work with probability as the fundamental quantity of uncertainty, we could use expectation
- de Finetti spent most of his life studying subjective conceptions of probability.
- He proposed the use of expectation as the primitive entity on which to base any analysis, as opposed to probability.
- Probabilities (where relevant) enter as derived quantities: they are the expectations of indicator functions.
- Note this asymmetry: if probability is treated as the primitive quantity then one has to specify (in the continuous case) an infinite set of probabilities in order to derive a single expectation.

Bayes linear

Bayes linear

Working with partial belief specifications

- In the Bayes linear approach, we follow de Finetti and take expectation as primitive.

Working with partial belief specifications

- In the Bayes linear approach, we follow de Finetti and take expectation as primitive.
- We construct partial belief specifications using only means, variances and covariances for all uncertain quantities

Working with partial belief specifications

- In the Bayes linear approach, we follow de Finetti and take expectation as primitive.
- We construct partial belief specifications using only means, variances and covariances for all uncertain quantities
- We may view the Bayes linear approach as
  - Offering a simple approximation to a full Bayes analysis
  - Complementary to the full Bayes approach, offering new interpretative and diagnostic tools
  - A generalisation of the full Bayes approach where we lift the restriction of requiring a full probabilistic prior before we may learn anything from data

Features of the Bayes linear approach

- Subjective and Bayesian

Features of the Bayes linear approach

- Subjective and Bayesian
- Belief specifications genuinely correspond to our beliefs

Features of the Bayes linear approach

- Subjective and Bayesian
- Belief specifications genuinely correspond to our beliefs
- Expectation as primitive

Features of the Bayes linear approach

- Subjective and Bayesian
- Belief specifications genuinely correspond to our beliefs
- Expectation as primitive
- Adjust beliefs by linear fitting rather than conditioning

Features of the Bayes linear approach

- Subjective and Bayesian
- Belief specifications genuinely correspond to our beliefs
- Expectation as primitive
- Adjust beliefs by linear fitting rather than conditioning
- Computationally straightforward allowing the analysis of larger and more complex problems

Features of the Bayes linear approach

- Subjective and Bayesian
- Belief specifications genuinely correspond to our beliefs
- Expectation as primitive
- Adjust beliefs by linear fitting rather than conditioning
- Computationally straightforward allowing the analysis of larger and more complex problems
- Diagnostic tools are an important part of the approach
    - How prior beliefs affect conclusions
    - How beliefs change by the adjustment
    - How beliefs about observables compare to the observations themselves

Features of the Bayes linear approach

- Subjective and Bayesian
- Belief specifications genuinely correspond to our beliefs
- Expectation as primitive
- Adjust beliefs by linear fitting rather than conditioning
- Computationally straightforward allowing the analysis of larger and more complex problems
- Diagnostic tools are an important part of the approach
  - How prior beliefs affect conclusions
  - How beliefs change by the adjustment
  - How beliefs about observables compare to the observations themselves
- Important special cases - multivariate Gaussian

Belief Specification

- The Bayes linear approach is subjectivist, and so in any analysis we need to specify our beliefs over all random quantities of interest.

Belief Specification

- The Bayes linear approach is subjectivist, and so in any analysis we need to specify our beliefs over all random quantities of interest.
- However, as we consider expectation as primitive we make our belief specifications in terms of the low-order moments of the random quantities of interest.

Belief Specification

- The Bayes linear approach is subjectivist, and so in any analysis we need to specify our beliefs over all random quantities of interest.
- However, as we consider expectation as primitive we make our belief specifications in terms of the low-order moments of the random quantities of interest.
- (If we have beliefs about higher orders we can include these in the analysis too)

Belief Specification

- The Bayes linear approach is subjectivist, and so in any analysis we need to specify our beliefs over all random quantities of interest.
- However, as we consider expectation as primitive we make our belief specifications in terms of the low-order moments of the random quantities of interest.
- (If we have beliefs about higher orders we can include these in the analysis too)
- For example, say we are interested in predicting $B = (B_1, B_2)^T$ from knowledge of $D = (D_1, D_2)^T$ which we will measure soon, then all we need to specify are $\mathrm{E}(B)$, $\mathrm{E}(D)$, $\mathrm{Var}(B)$, $\mathrm{Var}(D)$ and $\mathrm{Cov}(B, D)$.

The example: Numbers

- Suppose we have a four quantities of interest, $F = (B_1, B_2, D_1, D_2)^T$

The example: Numbers

- Suppose we have a four quantities of interest, $F = (B_1, B_2, D_1, D_2)^T$
- We observe values of $D = (D_1, D_2)^T$, and want to analyse the effects on our beliefs about B

The example: Numbers

- Suppose we have a four quantities of interest, $F = (B_1, B_2, D_1, D_2)^T$
- We observe values of $D = (D_1, D_2)^T$, and want to analyse the effects on our beliefs about B
- We have a very simple prior specification:

$$\mathrm{E}(F)_i = 0, \qquad \mathrm{Var}(F)_{ii} = 100,$$

and we have a correlation structure as follows

	B_1	B_2	D_1	D_2
B_1	1.00	0.56	0.52	0.61
B_2	0.56	1.00	0.32	0.98
D_1	0.52	0.32	1.00	0.28
D_2	0.61	0.98	0.28	1.00

Stages of belief analysis

A typical Bayes linear analysis of beliefs proceeds in the following stages:

1. Specification of prior beliefs
2. Interpret the expected adjustments a priori
3. Given observations, perform and interpret the adjustments
4. Make diagnostic comparisons between actual and expected beliefs

Bayes Linear Inference

Bayes Linear Inference

Belief Adjustment

- We are interested in how our beliefs about B change in the light of information given by D.

Belief Adjustment

- We are interested in how our beliefs about B change in the light of information given by D.
- We look among the collection of linear estimates, i.e. those of form $c_0 + c_1 D_1 + c_2 D_2$, and choose constants c_0, c_1, c_2 to minimise the prior expected squared error loss in estimating each of B_1 and B_2:

$$\mathrm{E}([B_1 - c_0 - c_1 D_1 - c_2 D_2]^2).$$

Belief Adjustment

- We are interested in how our beliefs about B change in the light of information given by D.
- We look among the collection of linear estimates, i.e. those of form $c_0 + c_1 D_1 + c_2 D_2$, and choose constants c_0, c_1, c_2 to minimise the prior expected squared error loss in estimating each of B_1 and B_2:
$$\mathrm{E}([B_1 - c_0 - c_1 D_1 - c_2 D_2]^2).$$
- The choices of constants may be easily computed, and the estimators $\mathrm{E}_D(B) = (\mathrm{E}_D(B_1), \mathrm{E}_D(B_2))^T$ turn out to be given by:
$$\mathrm{E}_D(B) = \mathrm{E}(B) + \mathrm{Cov}\,(B, D)\,\mathrm{Var}(D)^{\dagger}(D - \mathrm{E}(D)).$$
which we refer to as the adjusted expectation for collection B given collection D.

Adjusted expectation

- The adjusted expectation for collection B given collection D is

$$\mathrm{E}_D(B) = \mathrm{E}(B) + \mathrm{Cov}\,(B, D)\,\mathrm{Var}(D)^{\dagger}(D - \mathrm{E}(D)).$$

Adjusted expectation

- The adjusted expectation for collection B given collection D is

$$\mathrm{E}_D(B) = \mathrm{E}(B) + \mathrm{Cov}\,(B, D)\,\mathrm{Var}(D)^{\dagger}(D - \mathrm{E}(D)).$$

- The adjusted version of the B given D is the ‘residual’ vector

$$\mathbb{A}_B(D) = B - \mathrm{E}_D(B).$$

Adjusted expectation

- The adjusted expectation for collection B given collection D is

$$\mathrm{E}_D(B) = \mathrm{E}(B) + \mathrm{Cov}\,(B, D)\,\mathrm{Var}(D)^{\dagger}(D - \mathrm{E}(D)).$$

- The adjusted version of the B given D is the 'residual' vector

$$\mathbb{A}_B(D) = B - \mathrm{E}_D(B).$$

- We can partition the vector B as the sum of two uncorrelated vectors:

$$B = \mathrm{E}_D(B) + \mathbb{A}_B(D),$$

Adjusted variance

- We partition the variance matrix of B into two variance components:

$$\begin{aligned}\mathrm{Var}(B) &= \mathrm{Var}(\mathrm{E}_D(B)) + \mathrm{Var}(\mathbb{A}_B(D)) \\ &= \mathrm{RVar}_D(B) + \mathrm{Var}_D(B)\end{aligned}$$

Adjusted variance

- We partition the variance matrix of B into two variance components:

$$\begin{aligned}\mathrm{Var}(B) &= \mathrm{Var}(\mathrm{E}_D(B)) + \mathrm{Var}(\mathbb{A}_B(D)) \\ &= \mathrm{RVar}_D(B) + \mathrm{Var}_D(B)\end{aligned}$$

- These are the resolved variance matrix and the adjusted variance matrix (i.e. explained and residual variation).

Adjusted variance

- We partition the variance matrix of B into two variance components:

$$\begin{aligned}\mathrm{Var}(B) &= \mathrm{Var}(\mathrm{E}_D(B)) + \mathrm{Var}(\mathbb{A}_B(D)) \\ &= \mathrm{RVar}_D(B) + \mathrm{Var}_D(B)\end{aligned}$$

- These are the resolved variance matrix and the adjusted variance matrix (i.e. explained and residual variation).
- The variance matrices are calculated as

$$\begin{aligned}\mathrm{Var}_D(B) &= \mathrm{Var}(B) - \mathrm{Cov}\,(B, D)\,\mathrm{Var}(D)^{\dagger}\mathrm{Cov}\,(D, B)\,, \\ \mathrm{RVar}_D(B) &= \mathrm{Cov}\,(B, D)\,\mathrm{Var}(D)^{\dagger}\mathrm{Cov}\,(D, B)\,.\end{aligned}$$

Adjusted variance

- We partition the variance matrix of B into two variance components:

$$\begin{aligned}\mathrm{Var}(B) &= \mathrm{Var}(\mathrm{E}_D(B)) + \mathrm{Var}(\mathbb{A}_B(D)) \\ &= \mathrm{RVar}_D(B) + \mathrm{Var}_D(B)\end{aligned}$$

- These are the resolved variance matrix and the adjusted variance matrix (i.e. explained and residual variation).
- The variance matrices are calculated as

$$\mathrm{Var}_D(B) = \mathrm{Var}(B) - \mathrm{Cov}\left(B, D\right) \mathrm{Var}(D)^{\dagger} \mathrm{Cov}\left(D, B\right),$$

$$\mathrm{RVar}_D(B) = \mathrm{Cov}\left(B, D\right) \mathrm{Var}(D)^{\dagger} \mathrm{Cov}\left(D, B\right).$$

- Our variance matrices must be non-negative definite.

Adjusted variance

- We partition the variance matrix of B into two variance components:

$$\begin{aligned}\mathrm{Var}(B) &= \mathrm{Var}(\mathrm{E}_D(B)) + \mathrm{Var}(\mathbb{A}_B(D)) \\ &= \mathrm{RVar}_D(B) + \mathrm{Var}_D(B)\end{aligned}$$

- These are the resolved variance matrix and the adjusted variance matrix (i.e. explained and residual variation).
- The variance matrices are calculated as

$$\mathrm{Var}_D(B) = \mathrm{Var}(B) - \mathrm{Cov}(B, D)\,\mathrm{Var}(D)^\dagger \mathrm{Cov}(D, B),$$

$$\mathrm{RVar}_D(B) = \mathrm{Cov}(B, D)\,\mathrm{Var}(D)^\dagger \mathrm{Cov}(D, B).$$

- Our variance matrices must be non-negative definite.
- We use the Moore-Penrose generalized inverse ($A^\dagger$) to allow for degeneracy.

Resolution

- We summarize the expected effect of the data D for the adjustment of B by a scale-free measure which we call the resolution of B induced by D,

$$\mathrm{R}_D(B) = 1 - \frac{\mathrm{Var}_D(B)}{\mathrm{Var}(B)} = \frac{\mathrm{Var}(\mathrm{E}_D(B))}{\mathrm{Var}(B)}.$$

Resolution

- We summarize the expected effect of the data D for the adjustment of B by a scale-free measure which we call the resolution of B induced by D,

$$\mathrm{R}_D(B) = 1 - \frac{\mathrm{Var}_D(B)}{\mathrm{Var}(B)} = \frac{\mathrm{Var}(\mathrm{E}_D(B))}{\mathrm{Var}(B)}.$$

- The resolution lies between 0 and 1, and in general, small (large) resolutions imply that the information has little (much) linear predictive value, given the prior specification.

Resolution

- We summarize the expected effect of the data D for the adjustment of B by a scale-free measure which we call the resolution of B induced by D,

$$\mathrm{R}_D(B) = 1 - \frac{\mathrm{Var}_D(B)}{\mathrm{Var}(B)} = \frac{\mathrm{Var}(\mathrm{E}_D(B))}{\mathrm{Var}(B)}.$$

- The resolution lies between 0 and 1, and in general, small (large) resolutions imply that the information has little (much) linear predictive value, given the prior specification.
- Similar in spirit to an R^2 measure for the adjustment.

Example: The Adjustment

- We can calculate our adjusted expectations for points B given D algebraically as:

$$\mathrm{E}_D(B_1) = 0.381 D_1 + 0.507 D_2 + 0$$
$$\mathrm{E}_D(B_2) = 0.051 D_1 + 0.961 D_2 + 0$$

Example: The Adjustment

- We can calculate our adjusted expectations for points B given D algebraically as:

$$\mathrm{E}_D(B_1) = 0.381D_1 + 0.507D_2 + 0$$
$$\mathrm{E}_D(B_2) = 0.051D_1 + 0.961D_2 + 0$$

- We see that B_2 is mainly determined by the value of D_2 – unsurprising given the strength of $\mathrm{Corr}\left(B_2, D_2\right)$.

Example: The Adjustment

- We can calculate our adjusted expectations for points B given D algebraically as:

$$\mathrm{E}_D(B_1) = 0.381 D_1 + 0.507 D_2 + 0$$
$$\mathrm{E}_D(B_2) = 0.051 D_1 + 0.961 D_2 + 0$$

- We see that B_2 is mainly determined by the value of D_2 – unsurprising given the strength of $\mathrm{Corr}\,(B_2, D_2)$.
- We can also calculate the adjusted variance and resolutions

$$\mathrm{Var}_D(B) = \begin{pmatrix} 49.06 & -5.83 \\ -5.83 & 4.64 \end{pmatrix}, \quad \mathrm{R}_D(B) = \begin{pmatrix} 0.509 \\ 0.954 \end{pmatrix}$$

- We can see that we resolve much of the uncertainty about B_2

Example: Variance Partition

- We can decompose the prior variance into its resolved and unresolved portions:

$$\mathrm{Var}(B) = \mathrm{RVar}_D(B) + \mathrm{Var}_D(B)$$

$$\begin{pmatrix} 100.00 & 55.71 \\ 55.71 & 100 \end{pmatrix} = \begin{pmatrix} 50.94 & 61.54 \\ 61.54 & 95.36 \end{pmatrix} + \begin{pmatrix} 49.06 & -5.83 \\ -5.83 & 4.64 \end{pmatrix}$$

The observed adjustment

- Given the observed value d of D, we can calculate the observed adjusted expectation

$$\mathrm{E}_d(B) = \mathrm{E}(B) + \mathrm{Cov}\,(B, D)\,\mathrm{Var}(D)^{\dagger}(d - \mathrm{E}(D)).$$

The observed adjustment

- Given the observed value d of D, we can calculate the observed adjusted expectation

$$\mathrm{E}_d(B) = \mathrm{E}(B) + \mathrm{Cov}\,(B, D)\,\mathrm{Var}(D)^{\dagger}(d - \mathrm{E}(D)).$$

- For our example, we observe $d = (-8, 10)$ and the corresponding observed adjusted expectations are:

$$\mathrm{E}_d(B) = \begin{pmatrix} 2.02 \\ 9.20 \end{pmatrix}$$

- Having observed $D = d$, we notice that our adjusted expectations have both increased

The observed adjustment

- Given the observed value d of D, we can calculate the observed adjusted expectation

$$\mathrm{E}_d(B) = \mathrm{E}(B) + \mathrm{Cov}\,(B, D)\,\mathrm{Var}(D)^{\dagger}(d - \mathrm{E}(D)).$$

- For our example, we observe $d = (-8, 10)$ and the corresponding observed adjusted expectations are:

$$\mathrm{E}_d(B) = \begin{pmatrix} 2.02 \\ 9.20 \end{pmatrix}$$

- Having observed $D = d$, we notice that our adjusted expectations have both increased
- B_1 is weakly correlated with D and so is adjusted only a little, whereas B_2 is strongly correlated to D_2 and so its expectation shifts substantially towards the value $d_2 = 10$

Interpretation

Interpretation

Interpretations of belief adjustment

- An approximation
  - If we're fully Bayesian, then adjusted expectation is a tractable linear approximation to the full Bayes conditional expectation
  - Adjusted variance is then an easily-computable upper bound on the full Bayes preposterior risk, under quadratic loss

Interpretations of belief adjustment

- An approximation
  - If we're fully Bayesian, then adjusted expectation is a tractable linear approximation to the full Bayes conditional expectation
  - Adjusted variance is then an easily-computable upper bound on the full Bayes preposterior risk, under quadratic loss
- An estimator
  - $\mathrm{E}_D(B)$ is an 'estimator' of the value of B, which combines the data with simple aspects of our prior beliefs in a plausible manner
  - Adjusted variance is then the mean-squared error of the estimator $\mathrm{E}_D(B)$

Interpretations of belief adjustment

- An approximation
  - If we're fully Bayesian, then adjusted expectation is a tractable linear approximation to the full Bayes conditional expectation
  - Adjusted variance is then an easily-computable upper bound on the full Bayes preposterior risk, under quadratic loss
- An estimator
  - $\mathrm{E}_D(B)$ is an 'estimator' of the value of B, which combines the data with simple aspects of our prior beliefs in a plausible manner
  - Adjusted variance is then the mean-squared error of the estimator $\mathrm{E}_D(B)$
- A primitive
  - Adjusted expectation is a primitive quantification of further aspects of our beliefs about B having'accounted for' D
  - Adjusted variance is also a primitive, but applied to the 'residual variance' in B having removed the effects of D

Conditional Expectation

- The conditional expectation of $B|D$ is the value you would specify under the penalty $L_C = \sum_i cD_i[B - \mathrm{E}(B|D_i)]^2$

Conditional Expectation

- The conditional expectation of $B|D$ is the value you would specify under the penalty $L_C = \sum_i cD_i[B - \mathrm{E}(B|D_i)]^2$
- If D is a partition, so $D_i \in \{0, 1\}$ and $\sum_i D_i = 1$, then then the adjusted expectation minimises $L_A = \sum_i cD_i[B - x_i]^2$. So we choose x_i to be the conditional expectation, and

$$\mathrm{E}_D(B) = \sum_i \mathrm{E}(B|D_i)D_i$$

Conditional Expectation

- The conditional expectation of $B|D$ is the value you would specify under the penalty $L_C = \sum_i cD_i[B - \mathrm{E}(B|D_i)]^2$
- If D is a partition, so $D_i \in \{0, 1\}$ and $\sum_i D_i = 1$, then then the adjusted expectation minimises $L_A = \sum_i cD_i[B - x_i]^2$. So we choose x_i to be the conditional expectation, and

$$\mathrm{E}_D(B) = \sum_i \mathrm{E}(B|D_i)D_i$$

- So when D is a partition, the adjusted and conditional expectations are identical

Conditional Expectation

- The conditional expectation of $B|D$ is the value you would specify under the penalty $L_C = \sum_i cD_i[B - \mathrm{E}(B|D_i)]^2$
- If D is a partition, so $D_i \in \{0,1\}$ and $\sum_i D_i = 1$, then then the adjusted expectation minimises $L_A = \sum_i cD_i[B - x_i]^2$. So we choose x_i to be the conditional expectation, and

$$\mathrm{E}_D(B) = \sum_i \mathrm{E}(B|D_i)D_i$$

- So when D is a partition, the adjusted and conditional expectations are identical
- Adjusted expectation does not require D to be a partition, and so can be considered as a generalization of conditional expectation

Extension to linear combinations

- Let $\langle B \rangle$ be the set of all linear combinations of B

Extension to linear combinations

- Let $\langle B \rangle$ be the set of all linear combinations of B
- If $X = h^T B \in \langle B \rangle$, then we can write

$$\mathrm{E}(X) = h^T \mathrm{E}(B), \; \mathrm{Var}(X) = h^T \mathrm{Var}(B) h.$$

Extension to linear combinations

- Let $\langle B \rangle$ be the set of all linear combinations of B
- If $X = h^T B \in \langle B \rangle$, then we can write

$$\mathrm{E}(X) = h^T \mathrm{E}(B), \ \mathrm{Var}(X) = h^T \mathrm{Var}(B) h.$$

- So by specifying $\mathrm{E}(B)$ and $\mathrm{Var}(B)$ we have implicitly specified expectations and variances for all elements of $\langle B \rangle$

Extension to linear combinations

- Let $\langle B \rangle$ be the set of all linear combinations of B
- If $X = h^T B \in \langle B \rangle$, then we can write

$$\mathrm{E}(X) = h^T \mathrm{E}(B), \ \mathrm{Var}(X) = h^T \mathrm{Var}(B) h.$$

- So by specifying $\mathrm{E}(B)$ and $\mathrm{Var}(B)$ we have implicitly specified expectations and variances for all elements of $\langle B \rangle$
- Similarly, by calculating $\mathrm{E}_D(B)$ and $\mathrm{Var}_D(B)$, we have implicitly calculated the adjustment for all $X \in \langle B \rangle$

Diagnostics

Diagnostics

Data and Diagnostics

- Once data has been observed (first for D and then for B) we can perform diagnostics.

Data and Diagnostics

- Once data has been observed (first for D and then for B) we can perform diagnostics.
- The Bayes linear methodology has a rich variety of diagnostic tools available (more than in a fully Bayesian analysis).

Data and Diagnostics

- Once data has been observed (first for D and then for B) we can perform diagnostics.
- The Bayes linear methodology has a rich variety of diagnostic tools available (more than in a fully Bayesian analysis).
- We can perform diagnostics on individual random quantities, or on collections of random quantities.

Data and Diagnostics

- Once data has been observed (first for D and then for B) we can perform diagnostics.
- The Bayes linear methodology has a rich variety of diagnostic tools available (more than in a fully Bayesian analysis).
- We can perform diagnostics on individual random quantities, or on collections of random quantities.
- Three important versions are:
  - Prior Diagnostics.
  - Adjustment Diagnostics.
  - Final Observation Diagnostics.

Prior Diagnostics

- Each prior belief statement that we make describes our beliefs about some random quantity.

Prior Diagnostics

- Each prior belief statement that we make describes our beliefs about some random quantity.
- If we observe that quantity, we may compare what we expect to happen with what actually happens.

Prior Diagnostics

- Each prior belief statement that we make describes our beliefs about some random quantity.
- If we observe that quantity, we may compare what we expect to happen with what actually happens.
- Once we observe the values of $D = d$, we can check whether the data is consistent with our prior specifications.

Prior Diagnostics

- Each prior belief statement that we make describes our beliefs about some random quantity.
- If we observe that quantity, we may compare what we expect to happen with what actually happens.
- Once we observe the values of $D = d$, we can check whether the data is consistent with our prior specifications.
- For a single random quantity, we can calculate the standardized change and the discrepancy:

$$\mathrm{S}(d_i) = \frac{d_i - \mathrm{E}(D_i)}{\sqrt{\mathrm{Var}(D_i)}}, \quad \mathrm{Dis}(d) = \frac{[d_i - \mathrm{E}(D_i)]^2}{\mathrm{Var}(D_i)} = \mathrm{S}(d_i)^2$$

Prior Diagnostics

- Each prior belief statement that we make describes our beliefs about some random quantity.
- If we observe that quantity, we may compare what we expect to happen with what actually happens.
- Once we observe the values of $D = d$, we can check whether the data is consistent with our prior specifications.
- For a single random quantity, we can calculate the standardized change and the discrepancy:

$$\mathrm{S}(d_i) = \frac{d_i - \mathrm{E}(D_i)}{\sqrt{\mathrm{Var}(D_i)}}, \quad \mathrm{Dis}(d) = \frac{[d_i - \mathrm{E}(D_i)]^2}{\mathrm{Var}(D_i)} = \mathrm{S}(d_i)^2$$

- $\mathrm{E}(\mathrm{S}(d_i)) = 0$ and $\mathrm{Var}(\mathrm{S}(d_i)) = 1$, so if we observe $\mathrm{S}(d_i)$ greater than about 3 this suggests an inconsistency.

Discrepancy Ratio

- For the entire collection, the natural counterpart of the discrepancy is the Mahalanobis distance:

$$\mathrm{Dis}(d) = (d - \mathrm{E}(D))^T \mathrm{Var}(D)^\dagger (d - \mathrm{E}(D)).$$

Discrepancy Ratio

- For the entire collection, the natural counterpart of the discrepancy is the Mahalanobis distance:
$$\mathrm{Dis}(d) = (d - \mathrm{E}(D))^T \mathrm{Var}(D)^\dagger (d - \mathrm{E}(D)).$$
- The prior expected value of $\mathrm{Dis}(d)$ is given by $\mathrm{E}(\mathrm{Dis}(d)) = \mathbf{rk}\{\mathrm{Var}(D)\}$

Discrepancy Ratio

- For the entire collection, the natural counterpart of the discrepancy is the Mahalanobis distance:

$$\mathrm{Dis}(d) = (d - \mathrm{E}(D))^T \mathrm{Var}(D)^\dagger (d - \mathrm{E}(D)).$$

- The prior expected value of $\mathrm{Dis}(d)$ is given by $\mathrm{E}(\mathrm{Dis}(d)) = \mathbf{rk}\{\mathrm{Var}(D)\}$
- NB: if we pretend D is Normal, then $\mathrm{Dis}(d)$ would be χ^2

Discrepancy Ratio

- For the entire collection, the natural counterpart of the discrepancy is the Mahalanobis distance:

$$\mathrm{Dis}(d) = (d - \mathrm{E}(D))^T \mathrm{Var}(D)^\dagger (d - \mathrm{E}(D)).$$

- The prior expected value of $\mathrm{Dis}(d)$ is given by $\mathrm{E}(\mathrm{Dis}(d)) = \textbf{rk}\{\mathrm{Var}(D)\}$
- NB: if we pretend D is Normal, then $\mathrm{Dis}(d)$ would be χ^2
- We can then normalise the discrepancy, to obtain the discrepancy ratio for d

$$\mathrm{Dr}(d) = \frac{\mathrm{Dis}(d)}{\textbf{rk}\{\mathrm{Var}(D)\}},$$

which has prior expectation $\mathrm{E}(\mathrm{Dr}(d)) = 1$.

Discrepancy Ratio

- For the entire collection, the natural counterpart of the discrepancy is the Mahalanobis distance:

$$\mathrm{Dis}(d) = (d - \mathrm{E}(D))^T \mathrm{Var}(D)^\dagger (d - \mathrm{E}(D)).$$

- The prior expected value of $\mathrm{Dis}(d)$ is given by $\mathrm{E}(\mathrm{Dis}(d)) = \textbf{rk}\{\mathrm{Var}(D)\}$
- NB: if we pretend D is Normal, then $\mathrm{Dis}(d)$ would be χ^2
- We can then normalise the discrepancy, to obtain the discrepancy ratio for d

$$\mathrm{Dr}(d) = \frac{\mathrm{Dis}(d)}{\textbf{rk}\{\mathrm{Var}(D)\}},$$

  which has prior expectation $\mathrm{E}(\mathrm{Dr}(d)) = 1$.
- Large $\mathrm{Dr}(d)$ will of course also suggest inconsistencies.

Further Topics

Further Topics

Canonical Analysis

Canonical Analysis

Canonical analysis

- Our belief specification for B and our adjustment by D implies specifications and adjustments for all linear combinations in $\langle B \rangle$.

Canonical analysis

- Our belief specification for B and our adjustment by D implies specifications and adjustments for all linear combinations in $\langle B \rangle$.
- We can explore the (possibly complex) changes in beliefs about $\langle B \rangle$ induced by the adjustment via a canonical analysis

Canonical analysis

- Our belief specification for B and our adjustment by D implies specifications and adjustments for all linear combinations in $\langle B \rangle$.
- We can explore the (possibly complex) changes in beliefs about $\langle B \rangle$ induced by the adjustment via a canonical analysis
- A key component of the canonical analysis is the resolution transform matrix defined as

$$\mathbb{T}_{B:D} = \mathrm{Var}(B)^{\dagger}\mathrm{Cov}\,(B, D)\,\mathrm{Var}(D)^{\dagger}\mathrm{Cov}\,(D, B)\,.$$

Canonical analysis

- Our belief specification for B and our adjustment by D implies specifications and adjustments for all linear combinations in $\langle B \rangle$.
- We can explore the (possibly complex) changes in beliefs about $\langle B \rangle$ induced by the adjustment via a canonical analysis
- A key component of the canonical analysis is the resolution transform matrix defined as

$$\mathbb{T}_{B:D} = \mathrm{Var}(B)^{\dagger}\mathrm{Cov}\left(B, D\right)\mathrm{Var}(D)^{\dagger}\mathrm{Cov}\left(D, B\right).$$

- $\mathbb{T}_{B:D}$ has the property that $\mathrm{Var}(B)\mathbb{T}_{B:D} = \mathrm{RVar}_D(B)$

Canonical analysis

- Our belief specification for B and our adjustment by D implies specifications and adjustments for all linear combinations in $\langle B \rangle$.
- We can explore the (possibly complex) changes in beliefs about $\langle B \rangle$ induced by the adjustment via a canonical analysis
- A key component of the canonical analysis is the resolution transform matrix defined as

$$\mathbb{T}_{B:D} = \mathrm{Var}(B)^{\dagger}\mathrm{Cov}\,(B, D)\,\mathrm{Var}(D)^{\dagger}\mathrm{Cov}\,(D, B)\,.$$

- $\mathbb{T}_{B:D}$ has the property that $\mathrm{Var}(B)\mathbb{T}_{B:D} = \mathrm{RVar}_D(B)$
- The eigenstructure of $\mathbb{T}_{B:D}$ summarises all the effects of belief adjustment

Canonical analysis

- Our belief specification for B and our adjustment by D implies specifications and adjustments for all linear combinations in $\langle B \rangle$.
- We can explore the (possibly complex) changes in beliefs about $\langle B \rangle$ induced by the adjustment via a canonical analysis
- A key component of the canonical analysis is the resolution transform matrix defined as

$$\mathbb{T}_{B:D} = \mathrm{Var}(B)^{\dagger}\mathrm{Cov}\,(B, D)\,\mathrm{Var}(D)^{\dagger}\mathrm{Cov}\,(D, B)\,.$$

- $\mathbb{T}_{B:D}$ has the property that $\mathrm{Var}(B)\mathbb{T}_{B:D} = \mathrm{RVar}_D(B)$
- The eigenstructure of $\mathbb{T}_{B:D}$ summarises all the effects of belief adjustment
- Let the normed right eigenvectors of $\mathbb{T}_{B:D}$ be $v_1, \ldots, v_{r_B}$, ordered by eigenvalues $1 \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{r_B} \geq 0$ and scaled as $v_i^T \mathrm{Var}(B) v_i = 1$

Canonical directions

- We define the *i*th canonical direction as

$$Y_i = v_i^T(B - \mathrm{E}(B))$$

Canonical directions

- We define the ith canonical direction as
$$Y_i = v_i^T(B - \mathrm{E}(B))$$
- The canonical directions have the following properties
$$\mathrm{E}(Y_i) = 0, \quad \mathrm{Var}(Y_i) = 1, \quad \mathrm{Corr}\,(Y_i, Y_j) = 0$$
$$\mathrm{RVar}_D(Y_i) = \lambda_i, \quad \mathrm{Var}_D(Y_i) = 1 - \lambda_i,$$

Canonical directions

- We define the ith canonical direction as
$$Y_i = v_i^T (B - \mathrm{E}(B))$$
- The canonical directions have the following properties
$$\mathrm{E}(Y_i) = 0, \quad \mathrm{Var}(Y_i) = 1, \quad \mathrm{Corr}\,(Y_i, Y_j) = 0$$
$$\mathrm{RVar}_D(Y_i) = \lambda_i, \quad \mathrm{Var}_D(Y_i) = 1 - \lambda_i,$$
- So the collection $\{Y_1, Y_2, \ldots\}$ forms a mutually uncorrelated 'grid' of directions over $\langle B \rangle$, summarizing the effects of the adjustment.

Canonical directions

- We define the ith canonical direction as
$$Y_i = v_i^T(B - \mathrm{E}(B))$$
- The canonical directions have the following properties
$$\mathrm{E}(Y_i) = 0, \qquad \mathrm{Var}(Y_i) = 1, \qquad \mathrm{Corr}(Y_i, Y_j) = 0$$
$$\mathrm{RVar}_D(Y_i) = \lambda_i, \quad \mathrm{Var}_D(Y_i) = 1 - \lambda_i,$$
- So the collection $\{Y_1, Y_2, \ldots\}$ forms a mutually uncorrelated 'grid' of directions over $\langle B \rangle$, summarizing the effects of the adjustment.
- Y_1 is the quantity we learn most about. Y_2 is the quantity we learn next most about, given that it is uncorrelated with Y_1. $Y_{\mathbf{rk}\{B\}}$ is the quantity we learn least about.
- Relationship to canonical correlation analysis (and PCA)

Canonical properties and system resolution

- Each $X \in \langle B \rangle$ can be expressed using the canonical structure as

$$X - \mathrm{E}(X) = \sum_i \mathrm{Cov}\,(X, Y_i)\, Y_i,$$

$$\text{and } \mathrm{RVar}_D(X) = \sum_i \lambda_i (\mathrm{Corr}\,(X, Y_i))^2$$

Canonical properties and system resolution

- Each $X \in \langle B \rangle$ can be expressed using the canonical structure as

$$X - \mathrm{E}(X) = \sum_i \mathrm{Cov}\left(X, Y_i\right) Y_i,$$

$$\text{and } \mathrm{RVar}_D(X) = \sum_i \lambda_i (\mathrm{Corr}\left(X, Y_i\right))^2$$

- We can use this structure to express the resolved uncertainty for the entire collection $\langle B \rangle$ given adjustment by D via the resolved uncertainty and the system resolution

$$\mathrm{RU}_D(B) = \sum_i \lambda_i, \qquad \mathrm{R}_D(B) = \frac{1}{\mathbf{rk}\{B\}} \sum_i \lambda_i$$

Canonical properties and system resolution

- Each $X \in \langle B \rangle$ can be expressed using the canonical structure as

$$X - \mathrm{E}(X) = \sum_i \mathrm{Cov}\left(X, Y_i\right) Y_i,$$

$$\text{and } \mathrm{RVar}_D(X) = \sum_i \lambda_i (\mathrm{Corr}\left(X, Y_i\right))^2$$

- We can use this structure to express the resolved uncertainty for the entire collection $\langle B \rangle$ given adjustment by D via the resolved uncertainty and the system resolution

$$\mathrm{RU}_D(B) = \sum_i \lambda_i, \qquad \mathrm{R}_D(B) = \frac{1}{\mathbf{rk}\{B\}} \sum_i \lambda_i$$

- $\mathrm{R}_D(B)$ is a scalar summary of the effectiveness of the adjustment by D for the entire collection $\langle B \rangle$

Partial Analysis

Partial Analysis

Partial Analysis

- Suppose we have already adjusted out beliefs about B given data, D
  - Now suppose we get even more data F, how should we further adjust our beliefs about B?

Partial Analysis

- Suppose we have already adjusted out beliefs about B given data, D
  - Now suppose we get even more data F, how should we further adjust our beliefs about B?
- Suppose we have already adjusted our beliefs about B given data, $H = D \cup F$
  - What were the individual effects of adjusting by D or F?

Partial Analysis

- Suppose we have already adjusted out beliefs about B given data, D
  - Now suppose we get even more data F, how should we further adjust our beliefs about B?
- Suppose we have already adjusted our beliefs about B given data, $H = D \cup F$
  - What were the individual effects of adjusting by D or F?
- To answer either of these questions would require a partial analysis, where we consider the effects of subsets of the data on our beliefs

Partial adjustments

- If we adjust beliefs sequentially, then we can separate and scrutinize the adjustments at each stage

Partial adjustments

- If we adjust beliefs sequentially, then we can separate and scrutinize the adjustments at each stage
- We evaluate partial adjustments which represent the change in adjustment as we accumulate data.

Partial adjustments

- If we adjust beliefs sequentially, then we can separate and scrutinize the adjustments at each stage
- We evaluate partial adjustments which represent the change in adjustment as we accumulate data.
- Suppose we intend to adjust our beliefs about B by observations on D and F, we adjust B by $(D \cup F)$ but separate the effects of the subsets by adjusting B in stages, first by D, then adding F (or vice versa)

Partial adjustments

- If we adjust beliefs sequentially, then we can separate and scrutinize the adjustments at each stage
- We evaluate partial adjustments which represent the change in adjustment as we accumulate data.
- Suppose we intend to adjust our beliefs about B by observations on D and F, we adjust B by $(D \cup F)$ but separate the effects of the subsets by adjusting B in stages, first by D, then adding F (or vice versa)
- How do we separate the effects of D and F on B?

Separating things out

- If $D \perp\!\!\!\perp F$, then adjusted expectations are additive so

$$\mathrm{E}_{D \cup F}(B - \mathrm{E}(B)) = \mathrm{E}_D(B - \mathrm{E}(B)) + \mathrm{E}_F(B - \mathrm{E}(B))$$

Separating things out

- If $D \perp\!\!\!\perp F$, then adjusted expectations are additive so

$$\mathrm{E}_{D \cup F}(B - \mathrm{E}(B)) = \mathrm{E}_D(B - \mathrm{E}(B)) + \mathrm{E}_F(B - \mathrm{E}(B))$$

- If D and F are correlated, then we obtain a similar expression by removing the 'common variability' between F and D.

Separating things out

- If $D \perp\!\!\!\perp F$, then adjusted expectations are additive so

$$\mathrm{E}_{D\cup F}(B-\mathrm{E}(B)) = \mathrm{E}_D(B-\mathrm{E}(B)) + \mathrm{E}_F(B-\mathrm{E}(B))$$

- If D and F are correlated, then we obtain a similar expression by removing the 'common variability' between F and D.
- For any D, F, the vectors D and $\mathbb{A}_F(D) = F - \mathrm{E}_D(F)$ are uncorrelated.
- So, for any D, F

$$\mathrm{E}_{D\cup F}(B-\mathrm{E}(B)) = \mathrm{E}_D(B-\mathrm{E}(B)) + \mathrm{E}_{\mathbb{A}_F(D)}(B-\mathrm{E}(B))$$

The partial adjustment

- The partial adjustment of B by F given D, denoted $\mathrm{E}_{[F/D]}(B)$, is

$$\mathrm{E}_{[F/D]}(B) = \mathrm{E}_{D \cup F}(B) - \mathrm{E}_{D}(B) = \mathrm{E}_{\mathbb{A}_F(D)}(B - \mathrm{E}(B))$$

The partial adjustment

- The partial adjustment of B by F given D, denoted $\mathrm{E}_{[F/D]}(B)$, is

$$\mathrm{E}_{[F/D]}(B) = \mathrm{E}_{D \cup F}(B) - \mathrm{E}_D(B) = \mathrm{E}_{\mathbb{A}_F(D)}(B - \mathrm{E}(B))$$

- We can partition the variance in several ways

$$\begin{aligned} \mathrm{Var}(B) &= \mathrm{RVar}_D(B) + \mathrm{Var}_D(B) \\ &= \mathrm{RVar}_D(B) + \mathrm{RVar}_{[F/D]}(B) + \mathrm{Var}_{D \cup F}(B) \\ &= \mathrm{RVar}_{D \cup F}(B) + \mathrm{Var}_{D \cup F}(B) \end{aligned}$$

The partial adjustment

- The partial adjustment of B by F given D, denoted $\mathrm{E}_{[F/D]}(B)$, is

$$\mathrm{E}_{[F/D]}(B) = \mathrm{E}_{D \cup F}(B) - \mathrm{E}_D(B) = \mathrm{E}_{\mathbb{A}_F(D)}(B - \mathrm{E}(B))$$

- We can partition the variance in several ways

$$\begin{aligned}
\mathrm{Var}(B) &= \mathrm{RVar}_D(B) + \mathrm{Var}_D(B) \\
&= \mathrm{RVar}_D(B) + \mathrm{RVar}_{[F/D]}(B) + \mathrm{Var}_{D \cup F}(B) \\
&= \mathrm{RVar}_{D \cup F}(B) + \mathrm{Var}_{D \cup F}(B)
\end{aligned}$$

- The partial resolved variance matrix of B by F given D is

$$\mathrm{RVar}_{[F/D]}(B) = \mathrm{Var}(\mathrm{E}_{[F/D]}(B))$$

The end

The end

The end

We have seen:

- How we represent our beliefs – using expectation as primitive
- How we would update our beliefs – the BL adjustment
- How we can investigate potential problems in our belief specification – diagnostics
- How we can understand how our beliefs are affected by the data – canonical analysis
- How we would incorporate additional information – partial analysis