Modeling (Notes)

Mark DiBattista

January 17, 2020

Abstract

There are two parts to the modeling problem:

- Given a complex phenomenon or process, generate a finite set of structured datapoints, sampled at random or by design, that covers its range of properties or behavior;
- Given a finite set of structured datapoints, extract information that, given a new, partially complete datapoint, affords imputation of missing values from those present.

Usually, the information is incomplete and contradictory, and imputation provisional.

Models, and modeling techniques, are distinguished by the type of datapoints, the relationship among datapoints within a single sample, and linkage of datapoints across samples. The relationships within datapoints may be constrained in form, while linkage across samples maybe facilitated by sequencing or by mapping samples to an increasing variable such as time.

The focus of these notes is on the quantitative aspect of the modeling problem for which uncertainties in the approximating relationships and linkages are sufficiently regular to support the expression and interpretation of imputed values as estimates and confidence intervals.

1 Suggested Resource Materials

Useful source texts:

• Probability/Statistics, intermediate (probability sections are better than statistics):		*Statistical Inference*, Casella & Berger
• Probability, advanced:		*Probability and Measure*, Billingsley
• Computational Issues:	Linear Algebra	*Matrix Computations*, Golub & Van Loan
	: Collinearity	*Conditioning Diagnostics*, Belsley
• Generalized Linear Models:		*Multivariate Statistical Modelling Based on GLMs* Fahrmeir & Tutz

Throughout the text the acronyms refer to companion writeups,

LAN	*Linear Algebra (Notes)*
LAA	*Linear Algebra (Applications)*
PN	*Probability Notes*
SN	*Statistics Notes*

within which information is referenced by chapter and/or numbered equation.

2 Linear Models

The linear model covers the case for which the data points are real values, partitioned into a single response variable and remainder predictor variables, and the true relationship between predictors and response is a linear function. For any real problem the relationship encoded in the data points is noisy, however, and the assumption of linearity under common conditions for optimality is consistent with joint Gaussian distributions of response and predictor variables.

In fact given a set of measurements, there are two basic approaches commonly taken to solve the problem.

- The *engineering approach:* generate model coefficients for a prior linear relation between predictors and response variables that minimizes a loss function;
- The *probabilistic approach:* generate model parameters for a prior Gaussian conditional probability distribution, response variable conditioned on the predictors, that maximize an entropy measure.

The model parameters for which the loss function in the engineering approach is quadratic, also known as 'least squares', and for which the entropy measure is maximum likelihood are identical, and each can be shown to generate sample estimators for the mean vector and covariance matrix of the underlying joint Gaussian distribution. Finally, the estimated parameters of the Gaussian distribution can be used to generate interval tests and other measures of quality and stability of the model.

2.1 Gaussian Joint and Conditional Probability Function

A multivariate Gaussian distribution is completely defined by the specification of a mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$, with the probability density function

$$\mathbf{W} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow p_N(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})}. \tag{1}$$

The joint distribution can be expressed as the product of marginal and conditional distributions (*cf. PN*, §7.2.1.6) upon partitioning the variables into two distinct sets,

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix} \Rightarrow \begin{cases} \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \\ \\ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}, \end{pmatrix} \end{cases} \tag{2}$$

and the distribution of $\mathbf{W}_1$ conditioned on the realized variables, $\mathbf{W}_2 = \mathbf{w}_2$ is given by

$$\mathbf{W}_1 | (\mathbf{W}_2 = \mathbf{w}_2) \sim \mathrm{N}(\boldsymbol{\mu}_1 - (\mathbf{w}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}). \tag{3}$$

If the vector of random variables is partitioned into a single response variable, Y, conditioned on the remainder predictor variables, $X_1, \cdots, X_n$,

$$\mathbf{W} = \begin{pmatrix} Y \\ X_1 \\ \vdots \\ X_m \end{pmatrix} \Rightarrow \begin{cases} \boldsymbol{\mu} = \begin{pmatrix} \mu_y \\ \boldsymbol{\mu}_{\mathbf{x}} \end{pmatrix}, \boldsymbol{\mu}_{\mathbf{x}} = \begin{pmatrix} \mu_{x_1} \\ \vdots \\ \mu_{x_m} \end{pmatrix} \\ \\ \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{y\mathbf{x}}^\top \\ \boldsymbol{\sigma}_{y\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{xx}} \end{pmatrix}, \begin{cases} \boldsymbol{\sigma}_{y\mathbf{x}} = \begin{pmatrix} \sigma_{yx_1} & \cdots & \sigma_{yx_m} \end{pmatrix}^\top \\ \boldsymbol{\Sigma}_{\mathbf{xx}} = \begin{pmatrix} \sigma_{x_1}^2 & \cdots & \sigma_{x_1 x_m} \\ \vdots & \ddots & \vdots \\ \sigma_{x_1 x_m} & \cdots & \sigma_{x_m}^2 \end{pmatrix} \end{cases} \end{cases} \tag{4}$$

then the single-variate conditional distribution is also Gaussian, and takes the form,

$$Y | (\mathbf{X} = \mathbf{x}) \sim \mathrm{N}\left(\mu_y + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\sigma}_{y\mathbf{x}}, \sigma_y^2 - \boldsymbol{\sigma}_{y\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\sigma}_{y\mathbf{x}}\right). \tag{5}$$

The parameters of the Gaussian distribution can be simplified – and the linear form of the mean stressed – by introducing the following constants,

$$\left.\begin{array}{l} \mathbf{a} = \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} \\ \boldsymbol{\alpha} = \begin{pmatrix} \mu_y - \boldsymbol{\sigma}_{y\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\mu}_{\mathbf{x}} \end{pmatrix} \end{array}\right\} \Rightarrow \mu_{y|\mathbf{x}} = \mu_y + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\sigma}_{y\mathbf{x}} \equiv \boldsymbol{\alpha}^\top \mathbf{a} \tag{6}$$

$$R^2 = \frac{\boldsymbol{\sigma}_{y\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\sigma}_{y\mathbf{x}}}{\sigma_y^2} \qquad \Rightarrow \sigma_{y|\mathbf{x}}^2 = \sigma_y^2 \left(1 - R^2\right) \tag{7}$$

which leads to the expression,

$$Y | (\mathbf{X} = \mathbf{x}) \sim \mathrm{N}(\mu(\mathbf{x}), \sigma^2) \equiv \mathrm{N}\left(\boldsymbol{\alpha}^\top \mathbf{a}, \sigma_y^2 \left(1 - R^2\right)\right). \tag{8}$$

Here, the parameter, R^2, which appears in the variance of the conditional distribution, is the familiar coefficient of determination. It is clear from the derivation provided above, however, that the parameter describes the strength of the linear relationship between predictor and response variables, and is not an independent property of the measurement data.

2.2 Estimation of Linear Relationship from Data

The linear relationship between response and predictor variables is derived from information contained within the matrix, W, which contains a list of indexed measurements,

$$W = \begin{pmatrix} y_1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{nm} \end{pmatrix} = \begin{pmatrix} \mathbf{y} & X \end{pmatrix}. \tag{9}$$

Here, the measurements are organized by row, in which the first column contains the response variable, $\mathbf{y}$, and the remaining columns contain the predictors, X, and is called the **data matrix**. The predictor data points may be selected by design, or by chance.

We construct the **extended data matrix** by replacing the response vector with a unit vector,

$$A = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & X \end{pmatrix}, \tag{10}$$

which contains all information from which the linear relationship is to be derived. Each row in the extended data matrix in (10) matches the form of the vector of arbitrary predictor measurements in the conditional mean show above in (6).

$$\mathbf{a}_i^\top = \begin{pmatrix} 1 & x_{i1} & \cdots & x_{im} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_i^\top \end{pmatrix} \Rightarrow A = \begin{pmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_n^\top \end{pmatrix} \tag{11}$$

2.3 Engineering Approach: Least Squares

The engineering approach to linear modeling is quite direct: assert a linear relationship between the vector of response variables, $\mathbf{y}$, and the extended data matrix, A, in which the linear relationship is mediated through a vector of coefficients, $\boldsymbol{\alpha}$,

$$\mathbf{y} - A\boldsymbol{\alpha} = 0. \tag{12}$$

Typically, the number of data points, n, exceeds the number of predictors, m, and the linear relation in (12) is overdetermined. The properties of solutions to linear equations are covered in *LAN* §7.

Although there is generally no *exact* solution to the equation in (12), we can determine the *best* solution given a condition of optimality. For a quadratic penalty function the optimization problem leads to a gradient operator applied to an inner product:

$$\hat{\boldsymbol{\alpha}} \leftarrow \min_{\boldsymbol{\alpha}} ||\mathbf{y} - A\boldsymbol{\alpha}||_2^2 \Rightarrow \nabla_{\boldsymbol{\alpha}} \left(\mathbf{y} - A\boldsymbol{\alpha}\right)^\top \left(\mathbf{y} - A\boldsymbol{\alpha}\right) = 0. \tag{13}$$

This problem, covered also in *LAA* §3, is solved by expanding the inner product and applying the gradient operator term by term:

$$\begin{aligned} \nabla_{\boldsymbol{\alpha}} \left(\mathbf{y} - A\boldsymbol{\alpha}\right)^\top \left(\mathbf{y} - A\boldsymbol{\alpha}\right) &= \nabla_{\boldsymbol{\alpha}} \left(A\boldsymbol{\alpha}\right)^\top \left(A\boldsymbol{\alpha}\right) - \nabla_{\boldsymbol{\alpha}} \left(A\boldsymbol{\alpha}\right)^\top \mathbf{y} - \nabla_{\boldsymbol{\alpha}} \mathbf{y}^\top A\boldsymbol{\alpha} + \nabla_{\boldsymbol{\alpha}} \mathbf{y}^\top \mathbf{y} \\ &= \nabla_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top A^\top A\boldsymbol{\alpha} - \nabla_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top A^\top \mathbf{y} - \nabla_{\boldsymbol{\alpha}} \mathbf{y}^\top A\boldsymbol{\alpha} + \nabla_{\boldsymbol{\alpha}} \mathbf{y}^\top \mathbf{y} \\ &= 2A^\top A\hat{\boldsymbol{\alpha}} - 2A^\top \mathbf{y} = 0 \\ \Rightarrow \hat{\boldsymbol{\alpha}} &= \left(A^\top A\right)^{-1} A^\top \mathbf{y}. \end{aligned} \tag{14}$$

Notice that the least-squares solution is achieved by applying a linear operator derived from the extended data matrix, $(A^\top A)^{-1}A^\top$, to the predictor vector, $\mathbf{y}$. The solution is simply a linear combination of all measurements.

2.4 Probabilistic Approach: Maximum Entropy

The probabilistic approach to linear modeling asserts that each measurement is drawn from the same underlying distribution,

$$Y|(\mathbf{X} = \mathbf{x}) \sim \mathrm{N}\left(\boldsymbol{\alpha}^\top \mathbf{a}, \sigma^2\right) \Rightarrow p_{Y|\mathbf{X}=\mathbf{x}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(y - \boldsymbol{\alpha}^\top \mathbf{a}\right)^2} \tag{15}$$

for which the conditional distribution is taken as Gaussian, the linear relation is encoded in the mean, $\boldsymbol{\alpha}^\top \mathbf{a}$, and the strength of the relationship is encoded in the variance, σ^2. The determination of the coefficients to the linear model, $\boldsymbol{\alpha}$, is therefore point-estimation problem given sampled data, which is covered in *SN,* §3, while maximum-likelihood estimators are addressed specifically in §3.3.

If each sample is governed by the one-dimensional Gaussian distribution in (15), the full set of samples, assuming each is independent of the other, is governed by the multivariate Gaussian,

$$\left.\begin{aligned} \mathbf{Y} &= \begin{pmatrix} Y_1 & \cdots & Y_n \end{pmatrix}^\top \\ \mathbf{X} &= \begin{pmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_n \end{pmatrix}^\top \end{aligned}\right\} \Rightarrow \mathbf{Y}|(\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{pmatrix}^\top) = \prod_{i=1}^{n} \mathrm{N}\left(\boldsymbol{\alpha}^\top \mathbf{a}_i, \sigma^2\right) = \mathrm{N}(A\boldsymbol{\alpha}, \sigma^2 I_n). \tag{16}$$

Point estimators are derived from operations performed on the joint distribution; point estimators for the coefficients in the linear model, and the variance of the distribution are presented in the next few sections.

2.4.1 Conditional Mean

2.4.1.1 Maximum-Likelihood Estimator

The likelihood function for the sample distribution, which is also Shannon information, is simply the logarithm of the joint probability density function (*cf. PN,* §10.4),

$$p_{\mathbf{Y}|\mathbf{X}} = \frac{1}{\sigma^n\sqrt{(2\pi)^n}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-A\boldsymbol{\alpha})^\top(\mathbf{y}-A\boldsymbol{\alpha})}$$

$$\Rightarrow \ln p_{\mathbf{Y}|\mathbf{X}} = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-A\boldsymbol{\alpha})^\top(\mathbf{y}-A\boldsymbol{\alpha}) \quad (17)$$

Since the linear model coefficients enter the likelihood function as an inner product, the maximum-likelihood estimator is generated by the extrema of the gradient operator,

$$\nabla_{\boldsymbol{\alpha}}\left(\mathbf{y}-A\boldsymbol{\alpha}\right)^\top\left(\mathbf{y}-A\boldsymbol{\alpha}\right) = 0 \equiv \hat{\boldsymbol{\alpha}} \leftarrow \min_{\boldsymbol{\alpha}} ||\mathbf{y}-A\boldsymbol{\alpha}||_2^2 \Rightarrow \hat{\boldsymbol{\alpha}} = (A^\top A)^{-1}A^\top\mathbf{y}. \quad (18)$$

The optimization condition is identical to the least-squares problem shown above in §2.3, and so the solutions exactly coincide.

The probabilistic approach holds an advantage over the engineering approach, however, since the statistical parameters of the underlying joint distribution define a distribution for the parameters of the linear model. Indeed, given the mean and variance of the sample conditional distribution,

$$\mathbb{E}_{\mathbf{X}}\mathbf{Y} = A\boldsymbol{\alpha} \quad (19)$$

$$\mathbb{V}_{\mathbf{X}}\mathbf{Y} = \sigma^2 I \quad (20)$$

the mean and variance of the estimated coefficients, $\boldsymbol{\alpha}$, can be calculated directly,

$$\mathbb{E}_{\mathbf{X}}\hat{\boldsymbol{\alpha}} = \mathbb{E}_{\mathbf{X}}(A^\top A)^{-1}A^\top\mathbf{Y} = (A^\top A)^{-1}A^\top\mathbb{E}_{\mathbf{X}}\mathbf{Y} = (A^\top A)^{-1}A^\top A\boldsymbol{\alpha} = \boldsymbol{\alpha}, \quad (21)$$

$$\begin{aligned}\mathbb{V}_{\mathbf{X}}\hat{\boldsymbol{\alpha}} = \mathbb{V}_{\mathbf{X}}(A^\top A)^{-1}A^\top\mathbf{Y} = (A^\top A)^{-1}A^\top(\mathbb{V}_{\mathbf{X}}\mathbf{Y})A(A^\top A)^{-1} &= (A^\top A)^{-1}A^\top(\sigma^2 I)A(A^\top A)^{-1} \\ &= \sigma^2(A^\top A)^{-1}. \quad (22)\end{aligned}$$

Finally, since the underlying conditional distribution is Gaussian, and all arithmetic operations to generate the coefficients are linear, the *distribution* of the estimated coefficients must also be Gaussian (*cf. PN*, §7.2.1.4), and is given by

$$\hat{\boldsymbol{\alpha}} \sim \mathrm{N}\left(\boldsymbol{\alpha}, \sigma^2(A^\top A)^{-1}\right). \quad (23)$$

2.4.2 Conditional Variance

2.4.2.1 Maximum-Likelihood Estimator

The maximum-likelihood estimator for the variance of the conditional distribution, whose likelihood function is provided in (17), is the extremum of the derivative with respect to variance,

$$\begin{aligned}&\frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left(\mathbf{y}-A\boldsymbol{\alpha}\right)^\top\left(\mathbf{y}-A\boldsymbol{\alpha}\right)\right) = 0 \\ &\qquad\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\left(\mathbf{y}-A\hat{\boldsymbol{\alpha}}\right)^\top\left(\mathbf{y}-A\hat{\boldsymbol{\alpha}}\right) = \frac{1}{n}\left(\mathbf{y}-A(A^\top A)^{-1}A^\top\mathbf{y}\right)^\top\left(\mathbf{y}-A(A^\top A)^{-1}A^\top\mathbf{y}\right) \\ &\qquad\quad = \frac{1}{n}\left(\mathbf{y}-P_A\mathbf{y}\right)^\top\left(\mathbf{y}-P_A\mathbf{y}\right) = \frac{1}{n}\mathbf{y}^\top\left(I-P_A\right)\mathbf{y} \quad (24)\end{aligned}$$

The maximum-likelihood variance estimator is a quadratic form with the matrix operator, $I - P_A$, that projects vectors into the orthogonal complement of the column space of the extended data matrix, A.

2.4.2.2 Unbiased Estimator

The quadratic form in (24) shows that information in the variance is contained within the orthogonal complement of the column space of the data matrix, which is an $n - (m+1)$-dimensional subspace. Coupled with the observation that the mean is projected into the column space,

$$A\hat{\boldsymbol{\alpha}} = A(A^\top A)^{-1}A^\top \mathbf{y} = P_A\mathbf{y}, \tag{25}$$

which is an $(m+1)$-dimensional subspace, it is clear that the maximum-entropy estimators for the conditional Gaussian distribution generalize Fisher's Theorem (*cf. PN*, §7.2.1.7) to a multidimensional setting. The arguments that hold in the 1-dimensional case carry over to the multidimensional case, since

$$\left.\begin{aligned}\mathbb{E}_{\mathbf{X}}\left(\mathbf{Y}^\top\mathbf{Y}\right) &= n\sigma^2\\ \mathbb{E}_{\mathbf{X}}\left(\mathbf{Y}^\top P_A\mathbf{Y}\right) &= (m+1)\sigma^2\end{aligned}\right\} \Rightarrow \mathbb{E}_{\mathbf{X}}\left(\frac{1}{n-(m+1)}\mathbf{Y}^\top\left(I - P_A\right)\mathbf{Y}\right) = \sigma^2. \tag{26}$$

We can therefore adjust the normalizing factor in the maximum-likelihood variance estimator, which minimizes the overall average discrepancy between the estimated and actual variances, to derive a least-unbiased estimator,

$$\hat{s}^2 = \frac{1}{n-(m+1)}\mathbf{y}^\top\left(I - P_A\right)\mathbf{y}. \tag{27}$$

The distribution of the scaled unbiased estimator is then a chi-squared distribution with the appropriate degrees of freedom,

$$(n-(m+1))\frac{\hat{s}^2}{\sigma^2} \sim \chi^2_{n-(m+1)}. \tag{28}$$

2.4.2.3 Interval Tests for Linear Coefficient Estimators

Given that the

$$\frac{\mathbf{r}^\top(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})}{\sqrt{\hat{s}^2\mathbf{r}^\top(A^\top A)^{-1}\mathbf{r}}} \sim T(n-(m+1) \tag{29}$$

$$\mathbf{r} \equiv \mathbf{e}_i \Rightarrow \frac{\hat{\alpha}_i - \alpha_i}{\sqrt{\hat{s}^2(A^\top A)^{-1}_i}} \sim T(n-(m+1)) \tag{30}$$

2.4.3 Linear Operations of Least Squares and Maximum-Entropy Sample Statistics

The least-squares and maximum-entropy methods generate identical linear-algebraic operations and solutions. We show here that the arithmetic operations carried out in the matrix formulae encode detailed sample estimators for every mean, variance and covariance in the underlying joint distribution. The solutions are both equivalent to replacing all statistical quantities with sample estimates, and the matrix operations shown above are exactly those necessary to achieve this.

Given the extended data matrix, A, the **scatter matrix**, $A^\top A$, can be expressed as

$$A^\top A = \begin{pmatrix}\mathbf{1}^\top\\ X^\top\end{pmatrix}\begin{pmatrix}\mathbf{1} & X\end{pmatrix} = \begin{pmatrix}\mathbf{1}^\top\mathbf{1} & \mathbf{1}^\top X\\ X^\top\mathbf{1} & X^\top X.\end{pmatrix} \tag{31}$$

Each of these operations encode sample estimators for mean and covariance statistics (*cf. PN*, §3.3),

$$\begin{aligned} \mathbf{1}^\top \mathbf{1} &= n, && (32)\\ \mathbf{1}^\top X &= n\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top, && (33)\\ X^\top \mathbf{1} &= n\hat{\boldsymbol{\mu}}_{\mathbf{x}}, && (34)\\ X^\top X &= n\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}} + \hat{\boldsymbol{\mu}}_{\mathbf{x}}\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\right). && (35) \end{aligned}$$

The partitioned scatter matrix in (31) can be inverted by use of the lower Schur block matrix inversion formula (*cf. LAN*, §6.2), repeated here as (note the unrelated use of the variable, A),

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B\left(D - CA^{-1}B\right)^{-1}CA^{-1} & -A^{-1}B\left(D - CA^{-1}B\right)^{-1} \\ -\left(D - CA^{-1}B\right)^{-1}CA^{-1} & \left(D - CA^{-1}B\right)^{-1} \end{pmatrix}. \tag{36}$$

Upon assigning the sample statistics in the scatter matrix to the blocks in the Schur formula, we directly construct the inverse of the scatter matrix,

$$\left.\begin{aligned} A &= n \\ B &= n\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top \\ C &= n\hat{\boldsymbol{\mu}}_{\mathbf{x}} \\ D &= n\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}} + \hat{\boldsymbol{\mu}}_{\mathbf{x}}\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top\right) \end{aligned}\right\} \Rightarrow \left(A^\top A\right)^{-1} = \frac{1}{n}\begin{pmatrix} 1 + \hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{x}} & -\hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \\ -\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{x}} & \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \end{pmatrix} \tag{37}$$

Similarly, the means and covarances of the response variables are carried out by linear operation with the data matrix,

$$A^\top \mathbf{y} = \begin{pmatrix} \mathbf{1}^\top \mathbf{y} \\ X^\top \mathbf{y} \end{pmatrix} = n\begin{pmatrix} \hat{\mu}_y \\ \hat{\boldsymbol{\sigma}}_{y\mathbf{x}} + \hat{\mu}_y\hat{\boldsymbol{\mu}}_{\mathbf{x}} \end{pmatrix}. \tag{38}$$

Applying these results to the operations for calculating estimates of the conditional mean and variance

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= (A^\top A)^{-1}A^\top \mathbf{y} = \begin{pmatrix} \hat{\mu}_y - \hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\sigma}}_{y\mathbf{x}} \\ \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\sigma}}_{y\mathbf{x}} \end{pmatrix} && (39)\\ \hat{\sigma}^2 &= \frac{1}{n}\left(\mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top A(A^\top A)^{-1}A^\top \mathbf{y}\right) = \hat{\sigma}_y^2 - \hat{\boldsymbol{\sigma}}_{y\mathbf{x}}^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\sigma}}_{y\mathbf{x}} = \hat{\sigma}_y^2\left(1 - \hat{R}^2\right) && (40) \end{aligned}$$

we generate exactly the results achieved by replacing the statistical parameters in the conditional distribution in (8) with sample estimates.

Finally, the distribution of the coefficients of the linear model can be expressed as

$$\hat{\boldsymbol{\alpha}} \sim \mathrm{N}\left(\begin{pmatrix} \mu_y - \boldsymbol{\sigma}_{y\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_{\mathbf{x}} \end{pmatrix}, \frac{\sigma^2}{n}\begin{pmatrix} 1 + \boldsymbol{\mu}_{\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_{\mathbf{x}} & -\boldsymbol{\mu}_{\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \\ -\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_{\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \end{pmatrix}\right). \tag{41}$$

2.5 Sensitivity of the Linear Model

2.5.1 Condition Number

Much as the matrix condition number, (*cf. LAN,* §7.2), measures the stability of the solution to a linear equation in (*LAN,* (121)), the **least-squares condition number**, κ_{LS}, measures the stability of the overdetermined system above in (12). There are two sources of potential instability, including

- a dependency on the strength of the relationship between response and predictor variables, $\hat{R}$;

- a dependency on the condition number of the data matrix, $\kappa(A)$.

Given the solution to the linear model, $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{LS}$, and perturbations to the matrix and target vector, δA and $\delta\mathbf{y}$, respectively, the condition number, $\kappa_{LS}(A, \mathbf{y})$, is expressed as the relative magnitude of the perturbation to the solution:

$$\left.\begin{array}{l}\boldsymbol{\alpha}_{LS} \overset{\min}{\longleftarrow} A\boldsymbol{\alpha} = \mathbf{y}\\ (A+\delta A)(\boldsymbol{\alpha}+\delta\boldsymbol{\alpha}) = \mathbf{y}+\delta\mathbf{y}\end{array}\right\} \Rightarrow \frac{||\delta\boldsymbol{\alpha}||_2}{||\boldsymbol{\alpha}_{LS}||_2} \leq \kappa_{LS}(A, \mathbf{y}) \max\left(\frac{||\delta\mathbf{y}||_2}{||\mathbf{y}||_2}, \frac{||\delta A||_2}{||A||_2}\right). \tag{42}$$

In order to ensure non-singularity in the perturbed equation the relative magnitudes of the perturbations to matrix and target vector are bounded,

$$\epsilon = \max\left(\frac{||\delta A||_2}{||A||_2}, \frac{||\delta\mathbf{y}||_2}{||\mathbf{y}||_2}\right) < \frac{|\sigma_{\max}|}{|\sigma_{\min}|}, \tag{43}$$

and the least-squares condition number takes the form,

$$\kappa_{LS}(A, \mathbf{y}) = \kappa_2(A)(1 + \hat{R}^{-1}) + \kappa_2^2(A)(1 - \hat{R}^{-2})^{\frac{1}{2}}. \tag{44}$$

To derive the relation in (44) we parametrize the perturbed least-squares problem in terms of a continuous variable, t, for which the finite-level expression in (42) is captured at the point, $t = \epsilon$:

$$\left.\begin{array}{r}E = \epsilon^{-1}\delta A\\ \mathbf{f} = \epsilon^{-1}\delta\mathbf{y}\end{array}\right\} \Rightarrow (A+\delta A)(\boldsymbol{\alpha}+\delta\boldsymbol{\alpha}) = \mathbf{y}+\delta\mathbf{y} \Rightarrow \left.[(A+tE)\boldsymbol{\alpha}(t) = \mathbf{y}+tf]\right|_{t=\epsilon}, \tag{45}$$

and the perturbation to the solution is determined by the derivative to the parametrized equation,

$$\boldsymbol{\alpha}(t)|_{t=\epsilon} \approx \boldsymbol{\alpha}(0) + \epsilon\dot{\boldsymbol{\alpha}}(0) = \boldsymbol{\alpha}_{LS} + \delta\boldsymbol{\alpha}. \tag{46}$$

Taking the derivative of the parametrized equation and evaluating the result at the origin yields the result,

$$\begin{aligned}&\left.\frac{d}{dt}[(A+tE)^\top(A+tE)\boldsymbol{\alpha}(t) = (A+tE)^\top(\mathbf{y}+t\mathbf{f})]\right|_{t=\epsilon} \Rightarrow\\ &\qquad\qquad \dot{\boldsymbol{\alpha}}(0) = (A^\top A)^{-1}E^\top(\mathbf{y} - A\boldsymbol{\alpha}_{LS}) + (A^\top A)^{-1}A^\top(\mathbf{f} - E\boldsymbol{\alpha}_{LS}).\end{aligned} \tag{47}$$

Given the bounds

$$||E||_2 \leq ||A||_2 \tag{48}$$

$$||\mathbf{f}||_2 \leq ||\mathbf{y}||_2 \tag{49}$$

$$\frac{1}{||A||_2\,||\boldsymbol{\alpha}_{LS}||_2} \leq \frac{1}{||A\boldsymbol{\alpha}_{LS}||_2} = \frac{1}{||P_A\mathbf{y}||_2} \tag{50}$$

and the relations, which are easily established by expanding the matrices in terms of singular values, ((*cf. LAN,* §5.1), and the matrix condition number, (*cf. LAN,* §7.2),

$$||A||_2\,||(A^\top A)^{-1}A^\top||_2 = |\sigma_{\max}||\sigma_{\min}^{-1}| \quad = \kappa_2(A) \tag{51}$$

$$||A||_2^2\,||(A^\top A)^{-1}||_2 \quad = |\sigma_{\max}|^2|\sigma_{\min}^{-1}|^2 = \kappa_2^2(A) \tag{52}$$

we can take the norm of each relative term in the derivative:

$$\begin{aligned}&\frac{||(A^\top A)^{-1}||_2\,||E||_2\,||\mathbf{y} - A\boldsymbol{\alpha}_{LS}||_2}{||\boldsymbol{\alpha}_{LS}||_2} \leq \frac{||(A^\top A)^{-1}||_2\,||A||_2\,||(I - P_A)\mathbf{y}||_2}{||\boldsymbol{\alpha}_{LS}||_2} \leq\\ &\quad ||(A^\top A)^{-1}||_2\,||A||_2^2\,\frac{||(I - P_A)\mathbf{y}||_2}{||A||_2\,||\boldsymbol{\alpha}_{LS}||_2} \leq ||(A^\top A)^{-1}||_2\,||A||_2^2\,\frac{||(I - P_A)\mathbf{y}||_2}{||P_A\mathbf{y}||_2} = \kappa_2^2(A)(1 - \hat{R}^{-2})^{\frac{1}{2}},\end{aligned} \tag{53}$$

and

$$\frac{||(A^\top A)^{-1}A^\top||_2\,(||\mathbf{f}||_2 + ||E||_2\,||\boldsymbol{\alpha}_{LS}||_2)}{||\boldsymbol{\alpha}_{LS}||_2} \le ||(A^\top A)^{-1}A^\top||_2 \left(\frac{||\mathbf{y}||_2}{||\boldsymbol{\alpha}_{LS}||_2} + ||A||_2\right) \le$$
$$||(A^\top A)^{-1}A^\top||_2\,||A||_2 \left(\frac{||\mathbf{y}||_2}{||A||_2\,||\boldsymbol{\alpha}_{LS}||_2} + 1\right) \le ||(A^\top A)^{-1}A^\top||_2\,||A||_2 \left(\frac{||\mathbf{y}||_2}{||P_A\mathbf{y}||_2} + 1\right)$$
$$= \kappa_2(A)(1+\hat{R}^{-1}). \quad (54)$$

Substituting these values in the definition for the condition number yields the desired result,

$$\kappa_{LS}(A,\mathbf{y}) \equiv \frac{||\dot{\boldsymbol{\alpha}}(0)||_2}{||\boldsymbol{\alpha}(0)||_2} = \frac{||\delta\boldsymbol{\alpha}||_2}{||\boldsymbol{\alpha}_{LS}||_2} = \kappa_2(A)(1+\hat{R}^{-1}) + \kappa_2^2(A)(1-\hat{R}^{-2})^{\frac{1}{2}}. \quad (55)$$

Finally, note that the least-squares condition number is dominated by different powers of the matrix condition number for the two asymptotic bounds of the coefficient of determination, $\hat{R}$:

$$\hat{R} \to 1 \Rightarrow \kappa_{LS}(A,\mathbf{y}) \to 2\kappa_2(A) \quad (56)$$
$$\hat{R} \to 0 \Rightarrow \kappa_{LS}(A,\mathbf{y}) \to \kappa_2^2(A)\hat{R}^{-1}. \quad (57)$$

2.5.2 Collinearity Diagnostics

Although the least-squares condition number, described above in §2.5.1, provides a measure of the sensitivity of the solution to the linear model, it is insufficient to diagnose the specific causes of or specific contributions to instability. In particular it is insufficient to determine which variates might jointly contribute to instability through collinearity. In this section we present a collinearity diagnostic that

- provides clear thresholds for the presence of collinearity;
- identifies all variates that participate in a single collinear relationship;
- distinguishes between variate participation in multiple collinear relationships.

The matrix condition number, which figures strongly in the least-squares condition number, is affected both by the relative magnitudes and relative positions of the columns of the matrix, A. Collinearity, however, is a measure of the position of the column vectors alone, and the collinearity diagnostic operates on an adjusted matrix whose column vectors are scaled to unit magnitude:

$$\tilde{A} = AS \Rightarrow \tilde{\mathbf{a}}_i^\top \tilde{\mathbf{a}}_i = 1, \quad (58)$$

for which S is the appropriate diagonal matrix.

The collinearity diagnostics take two forms:

- the **condition indices**, which are the relative magnitude of the singular values of the scaled matrix:
$$\tilde{\eta}_i \equiv \frac{|\tilde{\sigma}_i|}{|\tilde{\sigma}_{\min}|} \quad (59)$$
- the **variance-decomposition** matrix, Π, for which $(i,j)^{th}$ entry, π_{ij}, is the normalized contribution of the j^{th} variate to the variance of the i^{th} coefficient, $(\mathbb{V}\tilde{\boldsymbol{\alpha}})_i \equiv (\tilde{A}^\top\tilde{A})_i^{-1}$. Using the Singular Value Decomposition, (*cf. LAN,* §5.1) for the scaled data matrix, $\tilde{A}$, we have
$$\tilde{A} = \tilde{U}\tilde{\Sigma}\tilde{V}^\top \Rightarrow (\tilde{A}^\top\tilde{A})^{-1} = \tilde{V}\tilde{\Sigma}^{-2}\tilde{V}^\top \Rightarrow \pi_{ij} = \frac{\tilde{v}_{ij}^2\tilde{\sigma}_i^{-2}}{\sum_j \tilde{v}_{ij}^2\tilde{\sigma}_j^{-2}}. \quad (60)$$

The presence of collinearity in the linear model is diagnosed by the following properties of condition indices and the variance-decomposition matrix:

- ‘large’ condition indices, together with
- two or more entries in the variance-decomposition matrix with at least one matching column index.

Depending on the the properties of data and model, the collinearity issues might be addressed by removing the appropriate contributors. Extensive discussion on diagnosing and correcting for collinearity, including detailed critiques of popular alternative methods, is provided in *Conditioning Diagnostics*.

2.6 Glossary of Statistical Terms Used in Linear Regression

Many presentations of linear regression are rooted in statistics, and employ standard statistical terms in describing problems, ranges and solutions.

Statistical Term	**Acronym**	**Pointwise Formula**	**Linear Algebraic Formula**	
Total Sum of Squares	TSS	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$\mathbf{y}^\top(I_n - P_{\mathbf{1}_n})\mathbf{y}$	(61)
Residual Sum of Squares	RSS	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$(\mathbf{y} - A\hat{\boldsymbol{\alpha}})^\top(\mathbf{y} - A\hat{\boldsymbol{\alpha}})$	(62)
Error Sum of Squares	SSE	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \equiv \sum_{i=1}^{n} e_i^2$	$\mathbf{y}^\top(I_n - P_A)\mathbf{y}$	(63)
Regression Sum of Squares	RSE	$\alpha_1 \sum_{i=1}^{n}(x_i - \bar{x}_i)^2$	$\alpha_1 \hat{\sigma}_x^2$	(64)
Residual Standard Error	RSE	$\sqrt{\frac{1}{n-(m+1)} \sum_{i=1}^{n} e_i^2}$	$\sqrt{\frac{1}{n-(m+1)} \mathbf{y}^\top(I_n - P_A)\mathbf{y}}$	(65)
Standard Error of Estimator	SE(α_i)	*depends...*	$\sqrt{\sigma^2 (A^\top A)_i^{-1}}$	(66)

$$\left.\begin{aligned} \hat{\alpha}_0 &= \bar{y} - \hat{\alpha}_1 \bar{x} \\ \hat{\alpha}_1 &= \tfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}\right\} \Rightarrow \hat{\boldsymbol{\alpha}} = (A^\top A)^{-1} A^\top \mathbf{y} = \begin{pmatrix} \hat{\mu}_y - \hat{\boldsymbol{\mu}}_{\mathbf{x}}^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \hat{\boldsymbol{\sigma}}_{y\mathbf{x}} \\ \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \hat{\boldsymbol{\sigma}}_{y\mathbf{x}} \end{pmatrix} \tag{67}$$

$$\left.\begin{aligned} \mathrm{SE}(\hat{\alpha}_0)^2 &= \sigma^2 \left(\tfrac{1}{n} + \tfrac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\ \mathrm{SE}(\hat{\alpha}_1)^2 &= \tfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}\right\} \Rightarrow \mathrm{diag}\left(\frac{\sigma^2}{n} \begin{pmatrix} 1 + \boldsymbol{\mu}_{\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\mu}_{\mathbf{x}} & -\boldsymbol{\mu}_{\mathbf{x}}^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \\ -\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\mu}_{\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \end{pmatrix}\right) \tag{68}$$

3 Linear Filters

A sequence of unobservable state vectors, $\mathbf{x}(t_i) \equiv \mathbf{x}_i$, evolving dynamically, is estimated from a matching set of measurements, $\mathbf{z}(t_i) \equiv \mathbf{z}_i$. The model is executed in two phases – evolution and measurement – each subject to characteristic error and uncertainty.

The state and measurement vectors may lie in different domains, but all functional relations are linear, with transitions or transformations mediated through matrices. And although the state may evolve continuously, information from measurements is received at discrete times, and we can model the process of evolution and measurement in two separate steps:

$$\text{Evolution: } \mathbf{x}_i = F_i\mathbf{x}_{i-1} + \mathbf{w}_i \tag{69}$$

$$\text{Measurement: } \mathbf{z}_i = H_i\mathbf{x}_i + \mathbf{v}_i \tag{70}$$

Here, the uncertainties in evolution and errors in measurement are captured by the noise vectors, $\mathbf{w}$ and $\mathbf{v}$, respectively.

The history of measurements, expressed as the realization of a sequence of random variables,

$$\mathbf{Z}_1^{i-1} \equiv (\mathbf{Z}_{i-1} = \mathbf{z}_{i-1}), \cdots, (\mathbf{Z}_1 = \mathbf{z}_1), \tag{71}$$

is the conditioning for the hidden state vector at time, t_i, at which time a new measurement is made, $\mathbf{Z}_i = \mathbf{z}_i$. The estimators for the hidden state vector can be modeled as the separate random variables,

$$\text{Evolution: } \mathbf{X}_{i|i-1} = \mathbf{X}_{i-1}|\mathbf{Z}_1^{i-1} \tag{72}$$

$$\text{Measurement: } \quad \mathbf{X}_i = \mathbf{X}_{i|i-1}|\mathbf{Z}_i, \tag{73}$$

each of which accounts for the additional information anticipated or received in the discrete two-step process described above in (69) and (70).

3.1 Kalman Filters

If the unobservable initial state vector, $\mathbf{x}_0$, is modeled by a Gaussian distribution, then all subsequent estimations of state take Gaussian form as well, since both evolution and measurement are executed as linear transformations. Therefore, both state and measurement vectors, $\mathbf{X}$ and $\mathbf{Z}$, can be imbedded in Gaussian probability models, assuming that

- The initial hidden state vector, $\mathbf{x}_0$, is modeled as Gaussian;
- the noise vectors, $\mathbf{w}$ and $\mathbf{v}$, are modeled as zero-mean Gaussians, independent both of each other and of themselves at different times:

$$\left.\begin{array}{r}\mathbf{W} \sim \mathrm{N}(0, Q)\\ \mathbf{V} \sim \mathrm{N}(0, R)\end{array}\right\} \Rightarrow \begin{cases}\mathbb{C}(Q_i, Q_j) = \mathbb{C}(R_i, R_j) = 0, i \neq j\\ \mathbb{C}(Q_i, R_j) = 0, i, j \in \mathbb{N}\end{cases} \tag{74}$$

Since all information in multidimensional Gaussian distributions is contained within the mean vectors and covariance matrices, the sequence of state changes is completely captured by the sequence of moments, which are linked by linear operations defined by the coupling matrices, F and H, and the noise covariance matrices, Q and R. The Kalman filter is by construction Markov, and the moments of a given iteration depend only on the moments of the prior iteration and the new information captured in the updated matrices and the new measurement. The Kalman filter identifies the most-probable state of the system with the mean of the Gaussian distribution, and the accuracy of the estimate with the covariance matrix.

3.1.1 Stages of the Kalman Filter

Although the underlying evolution of the state vector may evolve continuously, the Kalman filter is executed at discrete times in two stages, with one an evolution of the state, the other a measurement taken from the updated state. Given a prior state, the inferred posterior state and estimates of uncertainty of the inferred values are derived from the sequence of linear operations described below.

3.1.1.1 Prior Distribution

The hidden state variable is assumed to be well-modeled by a Gaussian distribution with known mean vector and covariance matrix,

$$\mathbf{X}_i \sim \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{75}$$

3.1.1.2 Evolution Stage

The evolution of the hidden state variable is assumed to follow the linear operation described above in (69),

$$\mathbf{X}_{i+1|i} = F_{i+1}\mathbf{X}_i + \mathbf{W}_{i+1}. \tag{76}$$

Linear operations on Gaussian distributions yield other Gaussian distributions whose moments are linear transformations of the original,

$$\boldsymbol{\mu}_{i+1|i} = \mathbb{E}\mathbf{X}_{i+1|i} = \mathbb{E}(F_{i+1}\mathbf{X}_i + \mathbf{W}_{i+1}) = F_{i+1}\boldsymbol{\mu}_i \tag{77}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{i+1|i} = \mathbb{V}\mathbf{X}_{i+1|i} &= \mathbb{E}\left(F_{i+1}\mathbf{X}_i + \mathbf{W}_{i+1} - \mathbb{E}(F_{i+1}\mathbf{X}_i + \mathbf{W}_{i+1})\right)\left(F_{i+1}\mathbf{X}_i + \mathbf{W}_{i+1} - \mathbb{E}(F_{i+1}\mathbf{X}_i + \mathbf{W}_{i+1})\right)^\top \\ &= F_{i+1}\left(\mathbb{E}(\mathbf{X}_i - \mathbb{E}\mathbf{X}_i)(\mathbf{X}_i - \mathbb{E}\mathbf{X}_i)^\top\right)F_{i+1}^\top + \mathbb{E}\mathbf{W}_{i+1}\mathbf{W}_i^\top \\ &= F_{i+1}\boldsymbol{\Sigma}_i F_{i+1}^\top + Q_{i+1}.\end{aligned} \tag{78}$$

The updated hidden state variable is modeled as the Gaussian distribution,

$$\mathbf{X}_{i+1|i} \sim \mathrm{N}(F_{i+1}\boldsymbol{\mu}_i, F_{i+1}\boldsymbol{\Sigma}_i F_{i+1}^\top + Q_{i+}) \equiv \mathrm{N}(\boldsymbol{\mu}_{i+1|i}, \boldsymbol{\Sigma}_{i+1|i}). \tag{79}$$

3.1.1.3 Measurement Stage

Although the distribution in (79) models the updated state variable, it is hidden from vew, and its values can only be inferred from measurement. Measurements are modeled by the linear operation described above in (70),

$$\mathbf{Z}_{i+1}|\mathbf{X}_{i+1|i} = H_{i+1}\mathbf{X}_{i+1|i} + \mathbf{V}_{i+1} \tag{80}$$

from which we derive the distribution of measured values,

$$\mathbf{Z}_{i+1}|\mathbf{X}_{i+1|i} \sim \mathrm{N}(H_{i+1}\boldsymbol{\mu}_{i+1|i}, H_{i+1}\boldsymbol{\Sigma}_{i+1|i}H_{i+1}^\top + R_{i+1}). \tag{81}$$

3.1.1.4 Posterior Distribution

Finally, the inferred posterior distribution for the updated state vector takes two pieces of information into account: the estimate for the evolution of state and the specific measurement made from the distribution that governs the updated state. Here, the posterior distribution can be expressed as a conditional distribution of the state vector conditioned on the measurement. The Kalman filter is a linear filter with all distributions Gaussian, assumptions that exactly match the linear model described above in §2. Therefore, we can imbed both the hidden state distribution and measurement distribution in a joint Gaussian distribution, and derive the posterior as the distribution of hidden state conditioned on the measurement.

A joint multidimensional Gaussian distribution can be partitioned in two blocks, as shown above in (2),

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Z}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mathbb{E}\mathbf{X}\\ \mathbb{E}\mathbf{Z}\end{pmatrix}, \begin{pmatrix}\mathbb{C}(\mathbf{X},\mathbf{X}) & \mathbb{C}(\mathbf{X},\mathbf{Z})\\ \mathbb{C}(\mathbf{Z},\mathbf{X}) & \mathbb{C}(\mathbf{Z},\mathbf{Z})\end{pmatrix}.\right) \tag{82}$$

Identifying the random variables, $\mathbf{X}$ and $\mathbf{Z}$, as representing the hidden state and measurement distributions, respectively, we can fill out the values in the mean vector and covariance matrix from the means and covariances defined in (79) and (81),

$$\left.\begin{array}{l}\mathbf{X} \rightarrow \mathbf{X}_{i+1|i}\\ \mathbf{Z} \rightarrow \mathbf{Z}_{i+1}|\mathbf{X}_{i+1}\end{array}\right\} \Rightarrow \left\{\begin{array}{l}\begin{pmatrix}\mathbb{E}\mathbf{X}\\ \mathbb{E}\mathbf{Z}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}_{i+1|i}\\ H_i\boldsymbol{\mu}_{i+1|i}\end{pmatrix}\\ \\ \begin{pmatrix}\mathbb{C}(\mathbf{X},\mathbf{X}) & \mathbb{C}(\mathbf{X},\mathbf{Z})\\ \mathbb{C}(\mathbf{Z},\mathbf{X}) & \mathbb{C}(\mathbf{Z},\mathbf{Z})\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{i+1|i} & \boldsymbol{\Sigma}_{i+1|i}H_i^\top\\ H_i\boldsymbol{\Sigma}_{i+1|i} & H_i\boldsymbol{\Sigma}_{i+1|i}H_i^\top + R_i\end{pmatrix}\end{array}\right. . \tag{83}$$

The conditional distribution is therefore given by (*cf. PN*, §7.2.1.6) and the equation in (3) above),

$$\mathbf{X}|(\mathbf{Z}=\mathbf{z}) \sim \mathrm{N}\left(\mathbb{E}\mathbf{X} + \mathbb{C}(\mathbf{X},\mathbf{Z})\mathbb{C}^{-1}(\mathbf{Z},\mathbf{Z})(\mathbf{z}-\mathbb{E}\mathbf{Z}), \mathbb{C}(\mathbf{X},\mathbf{X}) - \mathbb{C}(\mathbf{X},\mathbf{Z})\mathbb{C}^{-1}(\mathbf{Z},\mathbf{Z})\mathbb{C}(\mathbf{Z},\mathbf{X})\right). \tag{84}$$

For the Kalman filter it is convenient to define the 'gain matrix',

$$K(\mathbf{X},\mathbf{Z}) \equiv \mathbb{C}(\mathbf{X},\mathbf{Z})\mathbb{C}^{-1}(\mathbf{Z},\mathbf{Z}) \Rightarrow K_i \equiv \boldsymbol{\Sigma}_{i+1|i}H_i^\top(H_i\boldsymbol{\Sigma}_{i+1|i}H_i^\top + R_i)^{-1}, \tag{85}$$

so that the posterior distribution can be expressed as the Gaussian distribution,

$$\mathbf{X}_{i+1}|(\mathbf{Z}_{i+1}=\mathbf{z}_{i+1}) \equiv \mathbf{X}_{i+1} \sim \mathrm{N}\left(\boldsymbol{\mu}_{i+1|i} + K_i(\mathbf{z}_{i+1} - H_i\boldsymbol{\mu}_{i+1|i}), (I - K_iH_i)\boldsymbol{\Sigma}_{i+1|i}\right). \tag{86}$$

3.1.1.5 Summary of Iterated Estimates for Kalman Filter

The Kalman filter identifies the estimator for the hidden state as the mean vectors of the Gaussian distributions, and the measure of uncertainty of the estimate as the covariance matrices. We can summarize the actions required for execution by presenting the mean vectors and covariance matrices at each stage:

prior distribution

$$\hat{\mathbf{x}}_i = \mathbb{E}\mathbf{X}_i = \boldsymbol{\mu}_i \tag{87}$$

$$P_i = \mathbb{V}\mathbf{X}_i = \boldsymbol{\Sigma}_i \tag{88}$$

evolution distribution

$$\hat{\mathbf{x}}_{i+1|i} = \mathbb{E}\mathbf{X}_{i+1|i} = \boldsymbol{\mu}_{i+1|i} = F_{i+1}\boldsymbol{\mu}_i = F_{i+1}\hat{\mathbf{x}}_i \tag{89}$$

$$P_{i+1|i} = \mathbb{V}\mathbf{X}_{i+1|i} = F_{i+1}\boldsymbol{\Sigma}_iF_{i+1}^\top + Q_{i+1} = F_{i+1}P_iF_{i+1}^\top + Q_{i+1} \tag{90}$$

posterior distribution

$$\hat{\mathbf{x}}_{i+1} = \mathbb{E}\mathbf{X}_{i+1} = \boldsymbol{\mu}_{i+1|i} + K_{i+1}(\mathbf{z}_{i+1} - H_{i+1}\boldsymbol{\mu}_{i+1|i}) = \hat{\mathbf{x}}_{i+1|i} + K_{i+1}(\mathbf{z}_{i+1} - H_{i+1}\hat{\mathbf{x}}_{i+1|i}) \tag{91}$$

$$P_{i+1} = \mathbb{V}\mathbf{X}_{i+1} = (I - K_{i+1}H_{i+1})\boldsymbol{\Sigma}_{i+1|i} = (I - K_{i+1}H_{i+1})P_{i+1|i} \tag{92}$$

$$K_{i+1} = P_{i+1|i}H_{i+1}^\top(H_{i+1}P_{i+1|i}H_{i+1}^\top + R_{i+1})^{-1} \tag{93}$$

This can be neatly summarized in the table,

	Prior	**Intermediate**	**Posterior**
Estimator:	$\hat{\mathbf{x}}_i$	$\hat{\mathbf{x}}_{i+1\|i} = F_{i+1}\hat{\mathbf{x}}_i$	$\hat{\mathbf{x}}_{i+1} = \hat{\mathbf{x}}_{i+1\|i} + K_{i+1}(\mathbf{z}_{i+1} - H_{i+1}\hat{\mathbf{x}}_{i+1\|i})$
Uncertainty:	P_i	$P_{i+1\|i} = F_{i+1}P_iF_{i+1}^\top + Q_{i+1}$	$P_{i+1} = (I - K_{i+1}H_{i+1})P_{i+1\|i}$

Notice the role of the gain matrix, defined above in (85): the updated estimate for the hidden state vector is a weighted linear combination of the prior value, and the difference between measured and expected values, while the uncertainty that attends the update is progressively reduced as additional measurements are made. Both the weighting factors for combination and the reduction of uncertainty are controlled by the values of the gain matrix.

3.2 Linear Model as a Kalman Filter

We can recover the solution of the linear model, presented above in §2 and in *PAA*, §3.2, from the Kalman filter by introducing a trivial evolution equation, defined by

$$F_{i+1} = I; \tag{94}$$

$$Q_{i+1} = 0. \tag{95}$$

Furthermore, by changing names of the measurement variables,

$$H_{i+1} \to \mathbf{a}_{i+1}^\top \tag{96}$$

$$R_{i+1} \to 1 \tag{97}$$

$$z_{i+1} \to b_{i+1} \tag{98}$$

and the covariance and gain matrices,

$$P_i = (A_i^\top A_i)^{-1} \tag{99}$$

$$K_{i+1} \equiv \mathbf{k}_{i+1} = \frac{(A_i^\top A_i)^{-1}\mathbf{a}_{i+1}}{\mathbf{a}_{i+1}^\top (A_i^\top A_i)^{-1}\mathbf{a}_{i+1} + 1} \tag{100}$$

we find the iterative solution to the linear model as

$$\textbf{Estimator:} \qquad \hat{\mathbf{x}}_{i+1} = \left(I - \mathbf{k}_{i+1}\mathbf{a}_{i+1}^\top\right)\hat{\mathbf{x}}_i + \mathbf{k}_{i+1} b_{i+1} \tag{101}$$

$$\textbf{Uncertainty:} \qquad \left(A_{i+1}^\top A_{i+1}\right)^{-1} = \left(I - \mathbf{k}_{i+1}\mathbf{a}_{i+1}^\top\right)\left(A_i^\top A_i\right)^{-1} \tag{102}$$

4 Generalized Linear Models

The **linear model (LM)**, presented above in §2, possesses a few fundamental characteristics:

- The measured predictor and response data points are generated by a joint Gaussian distribution;
- The conditional distribution for the response variable is therefore Gaussian;
- The expectation of the response variable is linked to the predictors by a linear map;
- The 'model' is the conditional expectation of the response given predictor data.

The **generalized linear model (GLM)** is designed to model a wider class of problems by relaxing the constraints and extending the conditions placed on the LM:

- The joint distribution of response and predictor variables is not addressed;
- The conditional distribution for the response variable may be any member of the simple exponential family (*cf. PN,* §7.4);
- The expectation of the response variable is linked to a linear combination of the predictors by an arbitrary, user-specified map;
- The 'model' is the conditional expectation of the response given predictor data.

Calibration of both the LM and GLM is achieved by determining the coefficients of the linear combination of predictor variables that best fits the evidentiary data. However, since the joint distributional framework is left unaddressed and the linking function is within the user's control, the selection and calibration of the GLM is decidedly an exercise in engineering. Furthermore, the extension of allowable conditional distributions to members of the simple exponential family – for which the location parameters and response variable are coupled by simple multiplication – implies that distributional location parameters for independently generated ensembles of data points are coupled to the sample mean: the model is therefore derived, and the linear coefficients are determined, by maximizing ensemble likelihood (*cf. PN*, §11; *SN*, §3.3.3).

4.1 Derivation of the Generalized Linear Model

The simple exponential family of distributions for random vectors takes the form (*cf. PN*, §7.4 (234))

$$\mathbf{Y} \Rightarrow p(\mathbf{y}|\boldsymbol{\theta}) = h(\mathbf{y}) \exp(\boldsymbol{\theta}^\top \mathbf{y} - A(\boldsymbol{\theta})), \tag{103}$$

for which the coupling between distribution parameters and data point is expressed through the vector dot product. The GLM is based on a variation of the simple exponential family for which the distributional parameters are split into location and scale contributions,

$$\boldsymbol{\theta} = \begin{pmatrix} \xi \rightarrow \text{location parameter(s)} \\ \phi \rightarrow \text{scale parameter} \end{pmatrix}. \tag{104}$$

and the scale parameter is assumed to be known. The critical part of the distributions take exponential form in which the argument contains a term in which the dependent variable and location parameter are couple by simple multiplication, and a term that is a function of the location parameter alone,

$$Y \Rightarrow p(y|\boldsymbol{\theta}) \equiv p(y|\xi, \phi) \propto \exp\left(\frac{\xi y - b(\xi)}{\phi}\right). \tag{105}$$

As for all members of the exponential family of distributions, the mean and variance are expressed as derivatives of the log-partition function, $b(\xi)$, which is most easily demonstrated with the cumulant function (*cf. PN*, 5.1.3) . For the univariate case the moment-generating function takes the form,

$$\begin{aligned} M_y(u) = \mathbb{E}e^{uy} \propto & \int_{\mathbb{R}} \exp(uy) \exp\left(\frac{\xi y - b(\xi)}{\phi}\right) dy \\ &= \int_{\mathbb{R}} \exp\left(\frac{(\xi + \phi u)y - b(\xi + \phi u) + b(\xi + \phi u) - b(\xi)}{\phi}\right) dy \\ &= \exp\left(\frac{b(\xi + \phi u) - b(\xi)}{\phi}\right) \int_{\mathbb{R}} \exp\left(\frac{(\xi + \phi u)y - b(\xi + \phi u)}{\phi}\right) dy \\ &= \exp\left(\frac{b(\xi + \phi u) - b(\xi)}{\phi}\right). \end{aligned} \tag{106}$$

Since the cumulant function is the logarithm of the moment-generating function, and the mean and variance are the first two terms of the cumulant expansion, we have

$$K_y(u) = \ln M_y(u) = \frac{b(\xi + \phi u) - b(\xi)}{\phi} \Rightarrow \begin{cases} \mathbb{E}y = \frac{d}{du} K_y(u)\Big|_{u=0} & = b'(\xi) \\ \mathbb{V}y = \frac{d^2}{du^2} K_y(u)\Big|_{u=0} & = \phi b''(\xi) \end{cases}. \tag{107}$$

As with the LM, it is usually beneficial to augment the list of predictor variables for the GLM with a constant term, which ensures that the mean of the fit aligns with the data. Assigning the vector, $\mathbf{a}$, to the augmented data vector,

$$\mathbf{a} = \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix}, \tag{108}$$

the GLM requires that model predictions be constant on hyperplanes in the augmented space of predictor data,

$$\mathbf{a}^\top \boldsymbol{\alpha} \equiv \eta, \tag{109}$$

while calibration of the GLM is achieved by best orienting the direction of constant hyperplanes, defined by the vector of coefficients, $\boldsymbol{\alpha}$.

The relationship between mean value and the oriented hyperplanes is specified by the **link function**, g, although it more frequently appears in the presentation as the inverse map,

$$\eta = g(\mu) \Rightarrow \mu = g^{-1}(\eta). \tag{110}$$

From the definitions of low-order moments of the simple exponential family in (107) the mean can be immediately expressed in terms of the location parameter, ξ, through the derivative of the log-partition function, b, so that, for arbitrary link function, we can express the location parameter as a function of the predictor variables,

$$\mu(\xi) = b'(\xi) \Rightarrow \xi = \mu^{-1}\left(g^{-1}(\eta)\right). \tag{111}$$

Although the link function can be chosen arbitrarily to meet the needs of the model, the **natural link function** is the special case for which the link is identically the inverse mean function, $g = \mu^{-1}$. In this case the location parameter is exactly the linear combination of predictors, since we have

$$g = \mu^{-1} \Rightarrow \xi = \mu^{-1}\left(g^{-1}(\eta)\right) = \mu^{-1}\left(\mu\left(\eta\right)\right) = \eta \equiv \mathbf{a}^{\top}\boldsymbol{\alpha}. \tag{112}$$

Finally, the GLM identifies the expectation of the response variable, conditioned on the predictor variables, with the inverse of the user-specified link function,

$$\mathbb{E}Y = \mu = g^{-1}(\mathbf{a}^{\top}\boldsymbol{\alpha}). \tag{113}$$

4.2 Calibration of the Generalized Linear Model

For a specific datapoint, $\mathbf{x}_i$, the prediction provided by the GLM is calculated by linearly combining the augmented datapoint with the common coefficients, $\boldsymbol{\alpha}$, and applying the link function,

$$Y_i \Rightarrow p(y_i|\xi_i, \phi) \propto \exp\left(\frac{y_i\xi_i - b(\xi_i)}{\phi}\right) \Rightarrow \mathbb{E}Y_i = \mu_i = g^{-1}(\mathbf{a}_i^{\top}\boldsymbol{\alpha}). \tag{114}$$

For an IID ensemble of datapoints the GLM is calibrated using the collective information in the joint distribution,

$$\begin{aligned}\mathbf{Y} = \big(Y_1, \cdots, Y_n\big) \Rightarrow p(y_i, \cdots, y_n|\xi_1, \cdots, \xi_n, \phi) &\propto \prod_{i=1}^{N} \exp\left(\frac{y_i\xi_i - b(\xi_i)}{\phi}\right) \\ &= \exp\left(\frac{\sum_{i=1}^{n}\left(y_i\xi_i - b(\xi_i)\right)}{\phi}\right).\end{aligned} \tag{115}$$

Since the joint log-likelihood function for the simple exponential family reduces to the sum of individual log-likelihoods,

$$\ln L = \sum_{i=1}^{n} \ln L_i = \sum_{i=1}^{n} \ln p(y_i|\xi_i, \phi) \propto \sum_{i=1}^{n} \frac{y_i\xi_i - b(\xi_i)}{\phi}, \tag{116}$$

the maximum-likelihood estimator is determined by the vanishing sum of individual gradients,

$$\nabla_{\boldsymbol{\alpha}} \ln L = \sum_{i=1}^{n} \nabla_{\boldsymbol{\alpha}} \ln L_i = \sum_{i=1}^{n} \mathbf{S}_i = \mathbf{0}. \tag{117}$$

The key step to calculating the score function is handling the gradient of the log-partition function. From the inverse function theorem the derivative of the inverse is defined as follows,

$$(f^{-1})'(x) = (f'(f^{-1}(x)))^{-1}, \tag{118}$$

so that for members of the simple exponential family, whose mean and variance are related as in (107), we have

$$\left.\begin{aligned} f &\equiv b' = \mu \\ f' &\equiv b'' = \tfrac{\Sigma}{\phi} \end{aligned}\right\} \Rightarrow \left(\mu^{-1}\right)'\left(g^{-1}(\eta)\right) = \phi\Sigma^{-1}(\mu^{-1}(g^{-1}(\eta))). \tag{119}$$

The remainder of the effort is simply repeated applications of the chain rule,

$$\begin{aligned}\nabla_{\boldsymbol{\alpha}} \ln L_i = \nabla_{\boldsymbol{\alpha}} \frac{y_i \xi_i - b(\xi_i)}{\phi} = \frac{y_i - \mu_i}{\phi} \nabla_{\boldsymbol{\alpha}} \xi_i = \frac{y_i - \mu_i}{\phi} \nabla_{\boldsymbol{\alpha}} \mu^{-1}\left(g^{-1}(\eta_i)\right) \\ = (y_i - \mu_i)\left[\Sigma^{-1}(\mu^{-1}(g^{-1}(\eta_i)))\right]\left[(g^{-1})'(\eta_i)\right]\mathbf{a}_i. \quad (120)\end{aligned}$$

Here, the arguments can be suppressed – replaced by subscripts that indicate the dependence on the data points – to yield the simpler expression,

$$\nabla_{\boldsymbol{\alpha}} \ln L_i = (y_i - \mu_i)\ \Sigma_i^{-1}\,(g_i^{-1})'\,\mathbf{a}_i. \quad (121)$$

Finally, since GLMs are specific examples of maximum-likelihood estimators, the first moment of the score function vanishes – the mean residual is always zero – and the second moment is the Fisher information,

$$\begin{aligned}\mathbb{E}\mathbf{S}_i &= (\mathbb{E}y_i - \mu_i)\ \Sigma_i^{-1}\,(g_i^{-1})'\,\mathbf{a}_i = \mathbf{0} \quad (122)\\ \mathbb{V}\mathbf{S}_i &= \mathbb{E}\mathbf{S}_i\mathbf{S}_i^\top \equiv \mathbb{I}_F\mathbf{Y}_i = \mathbf{a}_i\,(g_i^{-1})'\,\Sigma_i^{-1}\,\mathbb{E}\,(y_i - \mu_i)^2\,\Sigma_i^{-1}\,(g_i^{-1})'\mathbf{a}_i^\top \\ &= (g_i^{-1})'\,\Sigma_i^{-1}\,(g_i^{-1})'\,\mathbf{a}_i\mathbf{a}_i^\top \quad (123)\end{aligned}$$

For the case in which the predictors and response mean are joined through the natural link function in (112) the structure simplifies considerably:

$$\text{natural link function: } \mu = g^{-1} \Rightarrow \begin{cases}\mathbf{S}_i = (y_i - \mu_i)\,\mathbf{a}_i \\ \\ \mathbb{E}\mathbf{S}_i = \mathbf{0} \\ \mathbb{V}\mathbf{S}_i = \Sigma_i\,\mathbf{a}_i\mathbf{a}_i^\top\end{cases} \quad (124)$$

Knowledge of he first and second moments of the GLM can be applied to two separate purposes:

- Given measured data, the estimated coefficients, $\hat{\boldsymbol{\alpha}}$, can be calibrated by a numerical method such as gradient descent, Newton's method, or other method from quadratic optimization;
- Given estimated coefficients, hypothesized values can be evaluated with statistical tests based on the asymptotic normality of maximum-likelihood estimators,

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \left(\sum_{i=1}^{n} \mathbb{I}_F\mathbf{Y}_i\right)^{-1}\right) = \mathrm{N}\left(\mathbf{0}, \left(\sum_{i=1}^{n} (g_i^{-1})'\,\Sigma_i^{-1}\,(g_i^{-1})'\,\mathbf{a}_i\mathbf{a}_i^\top\right)^{-1}\right). \quad (125)$$

4.3 Common Examples of Generalized Linear Models

4.3.1 Conditional Distribution: Bernoulli

Perhaps the best-known GLM arises from the case in which the response variables are independent boolean indicators and the predictor variables are assumed to act together through linear combination. The conditional distribution for the response variable is modeled through the Bernoulli distribution, which can be rearrange to satisfy the requirements of simple exponential form,

$$Y_i \sim \mathrm{Ber}(p_i) \Rightarrow \begin{cases} f(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i} \\ \qquad\qquad = \exp\left(y_i \ln\frac{p_i}{1-p_i} + \ln(1-p_i)\right)\end{cases}, y_i \in \{0,1\} \quad (126)$$

Here, the location parameter is the combination of variables against which the dependent variable is multiplied,

$$\xi_i = \ln\frac{p_i}{1-p_i} \Rightarrow \ln(1-p_i) = -\ln\left(1 + e^{\xi_i}\right), \quad (127)$$

and we can derive the mean of the distribution from the appropriate derivative of the log-partition function,

$$f(y_i|p_i) = \exp\left(y_i\xi_i - \ln\left(1+e^{\xi_i}\right)\right) \Rightarrow \mathbb{E}y_i = \frac{d}{d\xi_i}\ln\left(1+e^{\xi_i}\right) = \frac{e^{\xi_i}}{1+e^{\xi_i}} \equiv p_i. \tag{128}$$

The last equality is derived by inverting the relation in (127). We have recovered the obvious result: the parameter, p_i, is the mean of the Bernoulli distribution, which immediately yields the natural link function,

$$\Rightarrow p_i \equiv \mu_i = g^{-1}(\xi_i) = \frac{e^{\xi_i}}{1+e^{\xi_i}}. \tag{129}$$

There are a few common GLMs based on response variables modeled by Bernoulli distributions and conditioned on linear combinations of predictor variables. These common GLMs are differentiated by the user-specified link functions, which introduce the linearly combined predictor variables, η_i, into the model.

4.3.1.1 Logistic Regression

The natural link function for the Bernoulli distribution in (126) is the **logit function**,

$$\eta = g(\mu) = \ln\frac{\mu}{1-\mu} \Rightarrow \mu = g^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}}, \tag{130}$$

which can be inverted to provide the expected relation between location parameter and predictor variables,

$$\xi_i = \mu_i^{-1}\left(g^{-1}\left(\eta_i\right)\right) = \eta_i = \mathbf{a}_i^\top \boldsymbol{\alpha}. \tag{131}$$

The GLM is

$$\mathbb{E}Y_i = \hat{y}_i = \frac{\exp\left(\mathbf{a}_i^\top \hat{\boldsymbol{\alpha}}\right)}{1+\exp\left(\mathbf{a}_i^\top \hat{\boldsymbol{\alpha}}\right)}. \tag{132}$$

4.3.1.2 Probit Regression

5 Time Series

A **time series** is a list of random variables, $X_0, \cdots, X_n$, indexed by time, with the initial sample taken at time, t_0, and subsequent values taken at equal time increments, Δt,

$$X_0 = X(t_0), \tag{133}$$
$$X_i = X(t_0 + i\Delta t). \tag{134}$$

The various methods of time series analysis assert

- models for the uncertainty of random variables;
- relationships among random variables and uncertainties.

If the value of a random variable depends only on prior random variables, the time series is **causal**; if the distribution of the limiting random variable satisfies any finite bound, the time series is **non-divergent**; if the statistical properties of random variables, including means, variances and covariances, are constant in time, the time series is **stationary**.

A **measurement** is a realized value of a random variable, while an **innovation** is the difference between the realized and estimated values. Causal time series link measurements and innovations at one point in time to measurements and innovations at prior points.

Time series are examples of **discrete stochastic processes**.

5.1 ARIMA Process

If, given a causal time series, $X_0, \cdots, X_i, \cdots$, the model for uncertainty is constant and the relationship between any given random variable and prior random variables is fixed and linear, then we have an ARIMA process:

• AR:	Autoregressive	random variables are dependent on prior *measurements*;
• I:	Integrated	neighboring random variables are subject to differencing operations;
• MA:	Moving Average	random variables are dependent on prior *innovations*.

The number of prior measurements or innovations and the number of differencing operations are referred to as the **orders** of the process. A full ARIMA process is expressed as

$$\text{ARIMA}(p, d, q) \Rightarrow \begin{cases} \text{AR}(p) & \text{Random variables depend on } p \text{ prior measurements} \\ \text{I}(d) & \text{Random variables are subject to } d \text{ differencing operations} \\ \text{MA}(q) & \text{Random variables depend on } q \text{ prior innovations} \end{cases} \tag{135}$$

Most of the efforts here are directed towards AR processes, especially considering conditions under which the time series is stationary. We shall see that

- Non-stationary time series may be made stationary through differencing (integration);
- Autoregressive time series of finite order are equivalent to moving average time series of infinite order (and *v.v.*).

5.1.1 AR(p) Process

Given a time series of random variables, $X_0, \cdots, X_n$, an AR(p) model links each random variable to the p prior random variables by fixed linear combination, with each measurement subject to constant uncertainty,

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \epsilon_t \Rightarrow \begin{cases} \mathbb{E}\epsilon = 0 \\ \mathbb{V}\epsilon = 1 \end{cases} \quad p \le t \le n. \tag{136}$$

The equation in (136) can be more neatly expressed in terms of the backshift operator, $\mathcal{B}$, which decrements the index of a random variable,

$$\mathcal{B}X_t \equiv X_{t-1} \Rightarrow \left(1 - \sum_{j=1}^{p} \phi_j \mathcal{B}^j\right) X_t = \epsilon_t. \tag{137}$$

5.1.1.1 Formal Solution to an AR(p) Process

The backshift operator is linear, associative and distributive with respect to scalar multiplication – power expressions in back shift operators can therefore be factored and otherwise treated as polynomials. Indeed, we can factor the p^{th}-order backshift polynomial equation into p first-order terms, each governed by a root, λ_j, so that

$$1 - \sum_{j=1}^{p} \phi_j \mathcal{B}^j = \prod_{j=1}^{p} \left(1 - \lambda_j \mathcal{B}\right). \tag{138}$$

Upon application to random variables, the operators on both sides of the equation in (138) yield identical results.

Using the method of partial fractions, the inverse polynomial can be expanded in poles of the roots, so that

$$\left(1 - \sum_{j=1}^{p} \phi_j \mathcal{B}^j\right)^{-1} = \sum_{j=1}^{p} \gamma_j \left(1 - \lambda_j \mathcal{B}\right)^{-1}, \quad \gamma_j = \prod_{i \neq j} \frac{1}{\lambda_i - \lambda_j}. \tag{139}$$

The magnitude of the backshift operator is defined by its effects on random variables upon which it operates. Since the backshift operator changes only the index of the random variable with no effect on scale, the magnitude is unity. If any given root of the polynomial equation in (138) has magnitude less than unity, then the product of root and backshift operator is less than unity as well,

$$\left.\begin{array}{r} |\mathcal{B}| = 1 \\ |\lambda| < 1 \end{array}\right\} \Rightarrow |\lambda \mathcal{B}| < 1. \tag{140}$$

Finally, provided the magnitude of the root is less than unity, we can expand the poles in (139) into an infinite geometric series,

$$(1 - \lambda \mathcal{B})^{-1} = \sum_{k=0}^{\infty} \lambda^k \mathcal{B}^k. \tag{141}$$

Here, the equality in (141) follows from expanding the direct calculation,

$$(1 - \lambda \mathcal{B}) \sum_{k=0}^{\infty} \lambda^k \mathcal{B}^k = 1 + \sum_{k=1}^{\infty} \lambda^k \mathcal{B}^k - \lambda \mathcal{B} \sum_{k=0}^{\infty} \lambda^k \mathcal{B}^k = 1, \tag{142}$$

which holds provided the effects of operating on random variables encoded in the infinite sums converge. Convergence therefore requires the magnitude of the root be less than unity.

We can express the inverse operator in terms of the sums and products of powers and partial sums of roots,

$$\Gamma_k \equiv \Gamma_k(\lambda_1, \cdots, \lambda_p) = \sum_{j=1}^{p} \prod_{i \neq j}^{p} \frac{\lambda_j^k}{\lambda_i - \lambda_j}, \tag{143}$$

whose value is dominated by the largest root scaled by a constant. For the case in which all roots have magnitude less than unity, the elements vanish, dominated by a power series, so that

$$\lim_{k \to \infty} \Gamma_k = 0. \tag{144}$$

Applying the algebraic properties defined above, we can express the random variable, X_t, succinctly as an infinite series of backshift operations on the innovation, ϵ_t, provided the roots have magnitude all less than unity:

$$\begin{aligned}
X_t &= \left(1-\sum_{j=1}^{p}\phi_j\mathcal{B}^j\right)^{-1}\epsilon_t = \left(\sum_{j=1}^{p}\gamma_j\left(1-\lambda_j\mathcal{B}\right)^{-1}\right)\epsilon_t = \left(\sum_{j=1}^{p}\gamma_j\sum_{k=0}^{\infty}\lambda_j^k\mathcal{B}^k\right)\epsilon_t \\
&= \left(\sum_{j=1}^{p}\left(\prod_{i\neq j}^{p}\frac{1}{\lambda_i-\lambda_j}\right)\sum_{k=0}^{\infty}\lambda_j^k\mathcal{B}^k\right)\epsilon_t = \left(\sum_{k=0}^{\infty}\left(\sum_{j=1}^{p}\prod_{i\neq j}^{p}\frac{\lambda_j^k}{\lambda_i-\lambda_j}\right)\mathcal{B}^k\right)\epsilon_t \\
&\equiv \left(\sum_{k=0}^{\infty}\Gamma_k\mathcal{B}^k\right)\epsilon_t. \qquad (145)
\end{aligned}$$

Every stable finite-order AR model is formally equivalent to an infinite-order MA model.

5.1.1.2 Explicit Asymptotic Solution using Cyclic Permutation Matrices

We can calculate explicit solutions to the AR(p) process in (136) by identifying the shift operator, $\mathcal{B}$, applied to a time series of $n+1$ elements, with the $(n+1)\times(n+1)$ cyclic permutation matrix, U_{n+1}. This matrix has unit values on the super diagonal and in the lower left-hand corner. Repeated applications of the cyclic permutation matrix push the unit values onto progressively higher diagonals, terminating with the identity matrix:

$$\mathcal{B}\equiv U_{n+1} = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \end{pmatrix}, U_{n+1}^2 = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \end{pmatrix}, \cdots, U_{n+1}^n = I_{n+1} \qquad (146)$$

Properties of the shift operator are retained in the linear algebraic properties of cyclic permutation matrices:

- The inverse and transpose matrices coincide,

$$\left(U_{n+1}^k\right)^{-1} = \left(U_{n+1}^k\right)^{\top} \Rightarrow U_{n+1}^k\left(U_{n+1}^l\right)^{\top} = \begin{cases} U_{n+1}^{k-l}, & k>l \\ I_{n+1}, & k=l \\ U_{n+1}^{l-k}, & k<l \end{cases} \qquad (147)$$

- All eigenvalues, which can be calculated directly, have unit magnitude in any matrix norm,

$$\mathrm{eig}\left(U_{n+1}\right) = \left\{\exp\left(\frac{2\pi i}{n+1}j\right), j=1,\cdots,n+1\right\} \Rightarrow |\mathrm{eig}(U_{n+1})| = 1. \qquad (148)$$

The key ‘additional’ properties of cyclic permutation matrices, which introduce properties at variance with shift operators, are observed in (147): repeated applications of the inverse operation cycle through prior states, since $U_{n+1}^{n+k} = U_{n+1}^k$. This property is obviously non-causal; however, we show below that the error attributable to non-causal interactions is reduced as the number of observations, n, increases. In the limit of infinite observations the error vanishes, and causality is restored to the solution.

Interpreting the shift operator as a cyclic permutation matrix, an *augmented* version of the equation in (136) is given by,

$$\left(I_{n+1}-\sum_{j=1}^{p}\phi_j U_{n+1}^j\right)\boldsymbol{X} = \boldsymbol{\epsilon}, \qquad (149)$$

for which the sequence of random variables and innovations are expressed as vectors. Expanding the matrices into component form, the equation in (149) yields a banded set of linear equations that, for any *finite* n, has a unique solution. Note that the expression in (136) covers the top $n-p+1$ equations; the remainder 'wrap around' the interval and express the first p random variables partly as functions of the last p random variables.

$$\begin{pmatrix} 1 & -\phi_1 & \cdots & -\phi_p & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & -\phi_{p-1} & -\phi_p & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -\phi_1 & -\phi_2 & \cdots & -\phi_p & 0 \\ 0 & 0 & \cdots & 1 & 1 & -\phi_1 & \cdots & -\phi_{p-1} & -\phi_p \\ -\phi_p & 0 & \cdots & 0 & 0 & 1 & \cdots & -\phi_{p-2} & -\phi_{p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\phi_2 & -\phi_3 & \cdots & -\phi_p & 0 & 0 & \cdots & 1 & -\phi_1 \\ -\phi_1 & -\phi_2 & \cdots & -\phi_{p-1} & \phi_p & 0 & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} X_n \\ X_{n-1} \\ \vdots \\ X_{p+1} \\ X_p \\ X_{p-1} \\ \vdots \\ X_1 \\ X_0 \end{pmatrix} = \begin{pmatrix} \epsilon_n \\ \epsilon_{n-1} \\ \vdots \\ \epsilon_{p+1} \\ \epsilon_p \\ \epsilon_{p-1} \\ \vdots \\ \epsilon_1 \\ \epsilon_0 \end{pmatrix} \tag{150}$$

This is, of course, non-causal, but we shall show that the non-causal contribution to the solution decreases with increasing history, and vanishes completely in the asymptotic limit of infinite measurements.

The link between expressing the solution to the AR(p) process in terms of shift operators and cyclic permutation matrices is made by identifying both the identity and shift operator in (145) with the appropriate matrices,

$$\begin{aligned} \left.\begin{matrix} 1 \rightarrow I_{n+1} \\ \mathcal{B} \rightarrow U_{n+1} \end{matrix}\right\} &\Rightarrow \left(1 - \sum_{j=1}^{p} \phi_j \mathcal{B}^j\right) \rightarrow \left(I_{n+1} - \sum_{j=1}^{p} \phi_j U_{n+1}^j\right) \\ &\Rightarrow \sum_{j=1}^{p} \gamma_j \left(1 - \lambda_j \mathcal{B}\right)^{-1} \rightarrow \sum_{j=1}^{p} \gamma_j \left(I_{n+1} - \lambda_j U_{n+1}\right)^{-1} \\ &\Rightarrow \sum_{k=0}^{\infty} \Gamma_k \mathcal{B}^k \rightarrow \sum_{k=0}^{\infty} \Gamma_k U_{n+1}^k, \end{aligned} \tag{151}$$

which shows that, for the case in which all roots have magnitude less than unity, $\lambda_j < 1$, the solution takes the form,

$$\mathbf{X} = \left(\sum_{k=0}^{\infty} \Gamma_k U_{n+1}^k\right) \boldsymbol{\epsilon}. \tag{152}$$

For finite measurements, $n < \infty$, the solution can be further factored, since by the cyclic nature of the matrices,

$$\mathbf{X} = \left(\sum_{k=0}^{n} \left(\sum_{l=0}^{\infty} \Gamma_{kl}\right) U_{n+1}^k\right) \boldsymbol{\epsilon}. \tag{153}$$

Since by the relation in (144) the higher-order constants, dominated by a power series, vanish in the limit and we have

$$\lim_{n\rightarrow\infty} \sum_{k=0}^{n} \left(\sum_{l=0}^{\infty} \Gamma_{kl}\right) = \Gamma_k, \tag{154}$$

which implies that the non-causal contribution to the solution to the AR(p) process based on cyclic permutation matrices in (152) vanishes in the limit of infinite history.

5.1.1.3 Lagged Covariance Structure of the Formal Solution

The lagged autocovariance matrix of the solution in (152) can be constructed directly and, since the lagged autocovariance matrix for innovations is the identity, $\mathbb{E}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top = I_{n+1}$, takes the form,

$$\mathbb{E}\mathbf{X}\mathbf{X}^\top = \mathbb{E}\left(\left(\sum_{k=0}^{\infty}\Gamma_k U_{n+1}^k\right)\boldsymbol{\epsilon}\right)\left(\left(\sum_{l=0}^{\infty}\Gamma_l U_{n+1}^l\right)\boldsymbol{\epsilon}\right)^\top \tag{155}$$

$$= \left(\sum_{k=0}^{\infty}\Gamma_k U_{n+1}^k\right)\mathbb{E}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\left(\sum_{l=0}^{\infty}\Gamma_l U_{n+1}^l\right)^\top \tag{156}$$

$$= \left(\sum_{k=0}^{\infty}\Gamma_k U_{n+1}^k\right)\left(\sum_{l=0}^{\infty}\Gamma_l U_{n+1}^l\right)^\top. \tag{157}$$

Since the inverse and transpose matrices coincide for the cyclic permutation matrix, and nonzero entries in the summed matrices do not overlap, the entries in the asymptotic result as n diverges take the form,

$$\lim_{n\to\infty}\mathbb{E}\mathbf{X}\mathbf{X}^\top \equiv \boldsymbol{\Sigma} \Rightarrow \boldsymbol{\Sigma}_{ij} = \sum_{k=0}^{\infty}\Gamma_k\Gamma_{k+|i-j|}. \tag{158}$$

Furthermore, as the elements of the matrix depend only on the relative difference between i and j, the solution is stationary for constant coefficients, $\phi_1, \cdots, \phi_p$, provided the magnitudes of the roots, $\lambda_1, \cdots, \lambda_p$, are less than unity.

5.1.1.4 Parameter Estimation (Yule-Walker Equations)

The presentation in §§5.1.1.1 & 5.1.1.3 provides conditions under which the solution to the AR(p) processis stationary. Here, we take the complementary problem: given a set of data points, $x_0, \cdots, x_n$, estimate the parameters, $\phi_0, \cdots, \phi_p$, of the AR(p) process that best fits the data.

Taking the expectation of both sides of the equation in (136), we can express the AR(p) process as a linear combination of lagged autocovariances,

$$\mathbb{E}X_tX_{t-k} = \sum_{j-1}^{p}\phi_j\mathbb{E}X_tX_{t-k}. \tag{159}$$

Furthermore, assuming that the process is stationary,

$$\mathbb{E}X_tX_{t-k} = \mathbb{E}X_0X_k \tag{160}$$

we can form a linear equation – the **Yule-Walker equation** – whose coefficients are lagged autocovariances,

$$\begin{pmatrix} \mathbb{E}X_0X_0 & \mathbb{E}X_0X_1 & \cdots & \mathbb{E}X_0X_{p-1} \\ \mathbb{E}X_0X_1 & \mathbb{E}X_0X_0 & \cdots & \mathbb{E}X_0X_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}X_0X_{p-1} & \mathbb{E}X_0X_{p-2} & \cdots & \mathbb{E}X_0X_0 \end{pmatrix}\begin{pmatrix}\phi_1 \\ \phi_2 \\ \vdots \\ \phi_p\end{pmatrix} = \begin{pmatrix}\mathbb{E}X_0X_1 \\ \mathbb{E}X_0X_2 \\ \vdots \\ \mathbb{E}X_0X_p.\end{pmatrix} \tag{161}$$

Given the empirical lagged autocovariances,

$$\hat{\sigma}_{0k} = \frac{1}{n-k}\sum_{i=1}^{n-k}x_ix_{i+k} \tag{162}$$

the Yule-Walker equation can be inverted to yield the coefficients,

$$\left.\begin{aligned}\hat{\mathbf{\Sigma}} &= \begin{pmatrix} \hat{\sigma}_{00} & \hat{\sigma}_{01} & \cdots & \hat{\sigma}_{0,p-1} \\ \hat{\sigma}_{01} & \hat{\sigma}_{00} & \cdots & \hat{\sigma}_{0,p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{0,p-1} & \hat{\sigma}_{0,p-2} & \cdots & \hat{\sigma}_{00} \end{pmatrix} \\ \hat{\mathbf{v}} &= \begin{pmatrix} \hat{\sigma}_{01} \\ \hat{\sigma}_{02} \\ \vdots \\ \hat{\sigma}_{0p} \end{pmatrix} \end{aligned}\right\} \Rightarrow \boldsymbol{\phi} = \hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{v}}. \tag{163}$$

.

5.1.2 MA(q) Process

5.1.3 Unit Roots (I(d) Process)

The formal expression of the AR(p) model in terms of a polynomial expansion of the backshift operator, provided in (138), is given by

$$\left(1 - \sum_{j=1}^{p} \phi_j \mathcal{B}^j\right) X_t = \left(\prod_{j=1}^{p} (1 - \lambda_j \mathcal{B})\right) X_t = \epsilon_t. \tag{164}$$

As shown above, a solution to this equation exists provided that the magnitude of all roots be strictly less than unity, $|\lambda_j| < 1, 1 \leq j \leq p$.

If, however, there is a unit root of multiplicity, d, such that the factored polynomial can be rearrangeed as,

$$\prod_{j=1}^{p} (1 - \lambda_j \mathcal{B}) \equiv \left(\prod_{j=1}^{p-d} (1 - \lambda_j \mathcal{B})\right) (1 - \mathcal{B})^d, \tag{165}$$

for which the root indices are labeled so that the final d have unit value. Then the AR(p) process can be refactored as

$$\left(\prod_{j=1}^{p} (1 - \lambda_j \mathcal{B})\right) X_t = \left(\prod_{j=1}^{p-d} (1 - \lambda_j \mathcal{B})\right) (1 - \mathcal{B})^d X_t = \left(\prod_{j=1}^{p-d} (1 - \lambda_j \mathcal{B})\right) X_t^{'} \tag{166}$$

for which the unit root is incorporated directly into the random variable as a d^{th}-order differenced variable, $X_t^{'} = (1 - \mathcal{B})^d X_t$. Here, the transformed time series is stable, since the magnitude of all remaining roots are less than unity.

The absorption of the unit roots in (166) is equivalent to applying a first-order finite-difference scheme to the time series that approximates the d^{th} derivative operator.

5.1.3.1 Dickey-Fuller Test