# 8 The Contribution of Parameters to Stochastic Complexity

***Dean P. Foster and Robert A. Stine***
*Department of Statistics*
*The Wharton School of the University of Pennsylvania*
*Philadelphia, PA 19104-6302*
*foster@wharton.upenn.edu and stine@wharton.upenn.edu*

We consider the contribution of parameters to the stochastic complexity. The stochastic complexity of a class of models is the length of a universal, one-part code representing this class. It combines the length of the maximum likelihood code with the parametric complexity, a normalization that acts as a penalty against overfitting. For models with few parameters relative to sample size, $k \ll n$, the parametric complexity is approximately $\frac{k}{2} \log n$. The accuracy of this approximation, however, deteriorates as $k$ grows relative to $n$, as occurs in denoising, data mining, and machine learning. For these tasks, the contribution of parameters depends upon the complexity of the model class. Adding a parameter to a model class that already has many produces a different effect than adding one to a model class that has few. In denoising, for example, we show that the parametric complexity leads to an adaptive model selection criterion. We also address the calculation of the parametric complexity when the underlying integration is unbounded over the natural parameter space, as in Gaussian models.

## 8.1 Introduction, Terminology, and Notation

Parametric probability distributions $p_\theta$ provide a rich set of models for data compression, coding, and prediction. The parameters that distinguish these models often have clear physical ties to the underlying data, and so provide a comforting sense of reality and interpretation. The parameters can be linked to arrival rates, averages of underlying stochastic processes, or effects of exogenous influences that one seeks to control. When linked to a data-generating mechanism, both the number and values of the parameters $\theta$ take on substantive meaning that guides the choice of values for these tuning constants. When stripped of this connection and expanded in number, however, the choice of the best parameterization for $p_\theta$ becomes an alluring impediment. Modern computing makes it all too easy to expand the dimension of $\theta$ by adding superfluous parameters that promise much but deliver little. Indeed, overparameterized models that have been optimized to obtain the closest fit to data not only obscure any ties to an underlying data-generating mechanism but also predict poorly. Complex models found by automatic searches through massive data warehouses – data mining – nonetheless rule the day in modeling many phenomena. To choose one of these requires an automated criterion, and stochastic complexity stands out with appeal from many perspectives.

The routine use of stochastic complexity as a criterion to choose among complex models faces serious hurdles, however. These challenges arise in determining how to penalize for overparameterized models. Stochastic complexity appeared about 20 years ago [Rissanen 1986] and was found to possess a variety of optimality properties that spurred its use in hard problems. This optimality, though, lay in identifying parameters in models whose dimension remains fixed while the number of data records, $n$, expands. In data mining, the complexity of a model – reflected in the number of parameters – grows with the amount of data. The larger the data warehouse, the larger and more complex the variety of models one considers. If the dimension of $\theta$ grows with $n$, the standard asymptotic heuristics for stochastic complexity no longer obtain. For example, the familiar assessment of $\frac{1}{2}\log n$ per parameter no longer holds. Also, to make the procedure workable (in particular, to bound a key normalization), various artificial constraints have to be placed on the underlying probability models. These constraints can be provided in various forms with subtle implications for the choice of an optimal model.

We adopt the following notation and terminology that emphasize the connection between prefix codes and stochastic complexity. The response of interest is a sequence of $n$ values $\mathbf{y} = (y_1, \ldots, y_n)$, with each $y_i$ a point in some data space $D$ so that $\mathbf{y} \in D^n = D \times D \times \cdots \times D$. Our examples set $D$ to $\{0, 1\}$ for binary data and to the real line $\mathbb{R}$ in the Gaussian case. We assume that the space of possible outcomes $D$ is known. Rephrased as a problem in coding, the objective of model selection is to represent $\mathbf{y}$ using the shortest possible uniquely decodable prefix code. Here, “shortest possible” typically has one of two meanings. In a worst-case analysis, the chosen code for $\mathbf{y}$ is the solution of a minimax problem. Let $A$

denote a prefix-coding algorithm. For any $\mathbf{y} \in D^n$, the codebook associated with $A$ represents $\mathbf{y}$ using $\ell(A(\mathbf{y}))$ bits; an inverse lookup gives the decoding. A worst-case analysis seeks a code whose length attains the minimax rate

$$\min_{A} \max_{\mathbf{y} \in D^n} \ell(A(\mathbf{y})) \ . \tag{8.1}$$

Alternatively, one can define the best code as that with the shortest length with respect to some expectation [Barron, Rissanen, and Yu 1998].

The “models” that we study here are parametric probability distributions for the data, and so we will identify a specific codebook by its associated distribution. Because of the Kraft inequality, we can associate any prefix code with a (sub)probability distribution over $D^n$. Given a choice of parameters $\theta$ in some space $\Theta$, $p_\theta$ identifies the codebook for $\mathbf{y}$ implied by, say, arithmetic coding of $\mathbf{y}$ using the probability $p_\theta(\mathbf{y})$. Implicit in our notation is that one knows the form of the mapping that takes $\theta$ into a probability. For example, $p_{\mu,\sigma^2}$ could denote the normal distribution with mean $\mu$ and variance $\sigma^2$. One often collects a family of these models into classes, and here we use the term “library” for a collection of codebooks indexed by $\theta \in \Theta$,

$$\mathcal{L}(\Theta) = \{p_\theta : \theta \in \Theta\} \ . \tag{8.2}$$

Continuing with the Gaussian illustration, if $\Theta = \mathbb{R} \times \mathbb{R}^+$, then we have the independently and identically distributed (i.i.d.) Gaussian library

$$\mathcal{G}(\Theta) = \{p_{\mu,\sigma^2} : p_{\mu,\sigma^2}(\mathbf{y}) = \frac{e^{-\sum (y_i-\mu)^2/2}}{(2\pi\sigma^2)^{n/2}}, \mu \in \mathbb{R}, \sigma^2 > 0\} \ . \tag{8.3}$$

Calligraphic letters denote libraries; we use $\mathcal{L}$ to denote a generic library and use $\mathcal{B}$ and $\mathcal{G}$ for specific libraries. Notice that although any codebook $p_\theta$ identifies a prefix code for $\mathbf{y}$, a library $\mathcal{L}(\Theta)$ does not. We cannot encode $\mathbf{y}$ using $\mathcal{L}(\Theta)$ alone; either we must identify a specific $p_\theta \in \mathcal{L}(\Theta)$ or unify the library into a single codebook.

The following section defines stochastic complexity as the length of a prefix code for $\mathbf{y}$ obtained by an “encyclopedia,” a special codebook that represents a library. We introduce a special name for this codebook to distinguish it from the codebooks implied by parametric models $p_\theta$ that make up a library. With the terminology complete, Section 8.2 concludes with a guide to the rest of this chapter.

## 8.2 MDL and Stochastic Complexity

The *minimum description length* (MDL) criterion seeks the best library (model class) for encoding a particular sequence $\mathbf{y}$. The task is not to find the best individual codebook per se, but rather to identify a library. Since we assume that the mapping of parameters to codebooks $p_\theta$ has known form (given $\theta$), the problem becomes one of choosing the parameter space $\Theta$ rather than the form of

$p_\theta$. For example, we consider the problem of picking from regression models that are distinguished by the number of predictors rather than the comparison of linear regression to, say, other classes of generalized linear models.

To implement MDL thus requires a measure of how well a library can represent $\mathbf{y}$. Intuitively, one can proceed by first finding the maximum likelihood codebook in $\mathcal{L}(\Theta)$, say $p_{\hat{\theta}(\mathbf{y})}$. Since this codebook is indexed in a manner than depends upon $\mathbf{y}$, however, we cannot simply encode the data using the codebook $p_{\hat{\theta}(\mathbf{y})}$ alone because the receiver would not know which of the codebooks in $\mathcal{L}(\Theta)$ to use for the decoding. Two-part codes provide an obvious solution: identify the codebook in $\mathcal{L}(\Theta)$ by prefixing the code obtained by $p_{\hat{\theta}(\mathbf{y})}$ with another code identifying $\hat{\theta}(\mathbf{y})$. Through some clever arguments reviewed in [Rissanen 1989], Rissanen shows that one achieves a shorter overall code by coarsely identifying $\hat{\theta}(\mathbf{y})$. The use of two-part codes, however, introduces two problems. First, it is often neither easy nor obvious to decide how to round $\hat{\theta}(\mathbf{y})$; the discrete "spiral" codes given in [Rissanen 1983] illustrate some of the difficulties. Second, two-part codes are not "Kraft-tight"; the resulting implicit probability on $D^n$ sums to less than 1.

Stochastic complexity addresses both problems. First, it provides a direct construction that removes the subjective choice of how to encode $\hat{\theta}(\mathbf{y})$. Second, stochastic complexity encodes $\mathbf{y}$ with an efficient, one-part code. The underlying construction is rather natural: normalize the maximum likelihood code $p_{\hat{\theta}(\mathbf{y})}$ so that it becomes a probability. Since the data itself determine the maximum likelihood estimator (MLE), $p_{\hat{\theta}(\mathbf{y})}$ is not a subprobability,

$$\int_{D^n} p_{\hat{\theta}(\mathbf{y})}(\mathbf{y})d\mathbf{y} > 1 \ ,$$

(assuming a continuous model) and hence cannot define a prefix code for $\mathbf{y}$. The code length exceeds $\log 1/p_{\hat{\theta}}(\mathbf{y})$ in order to identify which codebook in $\mathcal{L}(\Theta)$ was used to represent the data. Rather than tack on a code that identifies $\hat{\theta}(\mathbf{y})$, one can instead convert the library back into a codebook. We distinguish these unified libraries from the parametric codebooks $p_\theta$ by calling them *encyclopedias*. The length of the code for $\mathbf{y}$ given by an encyclopedia is obtained by normalizing $p_{\hat{\theta}(\mathbf{y})}$ to generate a probability over $D^n$. This normalization requires us to divide by precisely the same integral that shows that $p_{\hat{\theta}(\mathbf{y})}$ is not a probability,

$$C\left(\mathcal{L}, \Theta, D^n\right) = \int_{D^n} p_{\hat{\theta}(\mathbf{y})}(\mathbf{y})d\mathbf{y} \ , \quad \text{where} \quad \hat{\theta}(\mathbf{y}) = \arg\max_{\theta\in\Theta} p_\theta(\mathbf{y}) \ . \tag{8.4}$$

Though this notation is cumbersome, we need these arguments to distinguish different forms of this normalization. The one-part code obtained from the resulting encyclopedia encodes $\mathbf{y}$ using the normalized maximum likelihood (NML) probability, denoted

$$g_{\mathcal{L}(\Theta)}(\mathbf{y}) = \frac{p_{\hat{\theta}(\mathbf{y})}(\mathbf{y})}{C\left(\mathcal{L}, \Theta, D^n\right)} \ . \tag{8.5}$$

The NML encyclopedia possesses many advantages. Not only can $g_{\mathcal{L}(\Theta)}$ be computed routinely without the need to round the MLE, the resulting Kraft-tight code obtains the minimax rate (8.1) [Shtarkov 1987].

The stochastic complexity of the library $\mathcal{L}(\Theta)$ for representing $\mathbf{y}$ is defined to be the length of the code provided by the resulting NML encyclopedia $g_{\mathcal{L}(\Theta)}$,

$$L(\mathbf{y}; \mathcal{L}, \Theta, D^n) = \log C(\mathcal{L}, \Theta, D^n) + \log \frac{1}{p_{\hat{\theta}(\mathbf{y})}(\mathbf{y})} \ , \quad \hat{\theta}(\mathbf{y}) \in \Theta \ . \tag{8.6}$$

The MDL criterion then picks the library that minimizes the stochastic complexity. The log of the normalizing constant, $\log C(\mathcal{L}, \Theta, D^n)$, is known as the *parametric complexity* of the library. It compensates for overfitting an excessive number of parameters; thus it acts like a penalty term.

The use of stochastic complexity can often be simplified by using a particularly simple asymptotic approximation for the parametric complexity. The underlying asymptotic analysis fixes the dimension of the parameter space $\Theta$ and lets the length $n$ tend to infinity. Under suitable regularity conditions, it follows that [Rissanen 1996]

$$\log C(\mathcal{L}, \Theta, D^n) = \frac{\dim(\Theta)}{2} \log \frac{n}{2\pi} + \log \int_{\Theta} |I(\theta)|^{1/2} d\theta + o(1), \tag{8.7}$$

where $I(\theta)$ is the asymptotic Fisher information matrix with elements

$$I_{ij}(\theta) = \lim_{n \to \infty} -\frac{1}{n} \frac{\partial^2 \log p_\theta(y)}{\partial \theta_i \partial \theta_j} \ . \tag{8.8}$$

The leading summand of (8.7) suggests that, in regular problems, the addition of each parameter increases the stochastic complexity by about $\frac{1}{2} \log n$. This interpretation motivates the common association of MDL with the Bayesian information criterion (BIC) whose penalty also grows logarithmically in $n$.

This approximation is both appealing and effective when used in the context of comparing a sequence of nested models of small dimension. For example, it works well in choosing among low-order polynomials or autoregressions (although comparisons tend to favor other criteria if prediction is the objective). For choosing among models of large dimension, such as those we use to predict credit risk [Foster and Stine 2002], however, the classic formulation of MDL (i.e., penalizing by the number of parameters times $\frac{1}{2} \log n$) no longer applies. For parameter-rich, data-mining models, this approximation no longer offers a useful measure of the complexity of the class.

The next three sections investigate the role of parameters in stochastic complexity, with an emphasis on models with many parameters. In Section 8.3, we consider the role of parameters in the Bernoulli library, a library that can be converted into an encyclopedia. We show that the contribution of a parameter depends on the complexity of the model itself; adding a parameter to a model with many adds less than adding one to a model with few. In Sections 8.4 and 8.5, we consider the parametric complexity of encyclopedias for Gaussian libraries. Section 8.4 considers methods for bounding the parametric complexity of a low-dimension Gaussian library, and

Section 8.5 considers high-dimensional models associated with denoising.

## 8.3 Parameters and the Bernoulli Library

We begin our discussion of stochastic complexity by choosing a context in which it all works. For this section, the data are binary with $D = \{0,1\}$. Codebooks for $\mathbf{y} \in \{0,1\}^n$ in the usual library $\mathcal{B}$ define probabilities of the form

$$p_\theta(\mathbf{y}) = \theta^{\sum y_i}\,(1-\theta)^{n-\sum y_i}, \tag{8.9}$$

with the parameter space $\Theta = [0,1]$. Given a binary sequence $\mathbf{y}$, the library $\mathcal{B}([0,1])$ of i.i.d. codebooks fixes $\theta$ for all $i$; larger parameter spaces allow this probability to vary over observations. In either case, we can compute the parametric complexity explicitly and see how the dimension of $\Theta$ affects the stochastic complexity.

The existence of a sufficient statistic simplifies this calculation. Under the assumed model class, the data are modeled as a realization of a sequence of independent Bernoulli random variables. Let $S_n = \sum_i Y_i$ denote the sum of these hypothetical random variables, and let $\hat\theta = S_n/n$ denote the MLE for $\theta$. The sufficiency of $S_n$ for $\theta$ allows us to factor the distribution of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ into the product of the distribution of $S_n$ and that of $\mathbf{Y}$ conditional on $S_n$ (which is thus free of $\theta$). Using these sufficiency arguments, the normalizing constant is

$$\begin{aligned} C(\mathcal{B},[0,1],\{0,1\}^n) &= \sum_{\mathbf{y}} p_{\hat\theta(\mathbf{y})}(\mathbf{y}) \\ &= \sum_{s=0}^{n} p_{\hat\theta(\mathbf{y})}(S_n = s) \sum_{\mathbf{y}:\hat\theta(\mathbf{y})=s/n} p(\mathbf{y} \mid S_n = s) \\ &= \sum_{s=0}^{n} p_{\hat\theta(\mathbf{y})}(S_n = s) \\ &= \sum_{s=0}^{n} \binom{n}{s} (s/n)^s (1-s/n)^{n-s}\ , \end{aligned} \tag{8.10}$$

where $p$ without a subscript denotes a probability distribution that is free of parameters. If we use Stirling's formula to approximate the factorials in (8.10), we obtain

$$\binom{n}{s}(s/n)^s(1-s/n)^{n-s} \approx \frac{\sqrt{n}}{\sqrt{2\pi s(n-s)}}\ .$$

This approximation is quite accurate except near the boundaries of the parameter space. (The approximation has singularities for $s = 0, n$, but the actual summands are 1. A Poisson approximation is more accurate at the extremes than this, essentially, normal approximation.) Integrating the approximation gives

$$C(\mathcal{B}([0,1]),[0,1],\{0,1\}^n) = \frac{\sqrt{n}}{\sqrt{2\pi}} \int_0^n \frac{1}{\sqrt{s(n-s)}} ds + O(1) = \sqrt{\frac{n\,\pi}{2}} + O(1)\ . \tag{8.11}$$

The error in this approximation is about 2/3.

The stochastic complexity (8.6) of the i.i.d. library $\mathcal{B}([0,1])$ for $\mathbf{y}$ is thus the sum of the code length for $\mathbf{y}$ plus the parametric complexity,

$$L(\mathbf{y}; \mathcal{B}, [0,1], \{0,1\}^n) = \tfrac{1}{2} \log \tfrac{n\pi}{2} + \log 1/p_{\hat{\theta}(\mathbf{y})}(\mathbf{y}) + O(1/\sqrt{n}) \ .$$

The parametric complexity agrees with the asymptotic approximation (8.7). The one parameter $\theta$ contributes about $\frac{1}{2}\log n$ the stochastic complexity.

The stochastic complexity of $\mathcal{B}(\Theta)$ is invariant of one-to-one transformations of $\Theta$, even if such a transformation makes $\Theta$ unbounded. For example, if we write $p_\theta$ in the canonical form of an exponential family, then

$$p_\theta(y) = e^{y \log \theta/(1-\theta) + \log 1-\theta}, \quad y = 0, 1,$$

or

$$p_\eta(y) = e^{y\,\eta + \psi(\eta)}, \quad y = 0, 1,$$

with $\eta = \log \theta/(1-\theta)$, the log of the odds ratio. Expressed in this form, the parameter space becomes $\mathbb{R}$. The stochastic complexity remains the same, though, since transforming the parameter space does not change the likelihood obtained by the various codebooks. The MLE for $\eta$ is $\hat{\eta} = \log \hat{\theta}/(1-\hat{\theta})$ and

$$\sum_{\mathbf{y}} p_{\hat{\eta}(\mathbf{y})}(\mathbf{y}) = \sum_{\mathbf{y}} p_{\hat{\theta}(\mathbf{y})}(\mathbf{y}) \ .$$

The contribution of a parameter does change, however, if we expand $\Theta$ to dimensions on the order of the number of observations. While artificial, perhaps, in this context, the use of stochastic complexity in data mining requires one to assess and compare models of large dimension. With a richer class of models, we no longer obtain an appealing separation of parameters from data. In such problems, the asymptotic approximation (8.7) fails because the dimension of $\Theta$ grows with $n$. An alternative, local asymptotic analysis leads to a rather different characterization of the amount to penalize for each parameter, one for which the penalty is proportional to the number of parameters rather than $\log n$ [Foster and Stine 1999].

Consider the "saturated" Bernoulli library $\mathcal{B}$ with the parameter space extended to $\Theta = [0,1] \times [0,1] \times \cdots \times [0,1] = [0,1]^n$, allowing one parameter for each observation. The MLE for $\theta^n = (\theta_1, \ldots, \theta_n)$ is $\hat{\theta}^n(\mathbf{y}) = \mathbf{y}$. As a result, the length of the maximum likelihood code for $\mathbf{y}$ collapses to zero,

$$\log \frac{1}{p_{\hat{\theta}^n(\mathbf{y})}(\mathbf{y})} = \log 1 = 0 \ .$$

The parametric complexity of $\mathcal{B}([0,1]^n)$ now comprises *all* of the stochastic complexity of the encyclopedia, with all of the information from the data concentrated in the parameters,

$$\log C(\mathcal{B},[0,1]^n,\{0,1\}^n) = \log \sum_{\mathbf{y}} p_{\hat{\theta}^n(\mathbf{y})}(\mathbf{y}) = \log 2^n = n \ .$$

Each parameter contributes just 1 bit, not $\frac{1}{2}\log n$, to the complexity of $\mathcal{B}([0,1]^n)$. Parameters in libraries for which the dimension of $\Theta$ is $O(n)$ evidently add less to the complexity than those in models of small, fixed dimension.

The concentration of the stochastic complexity into the parametric complexity leads to a dilemma when one then tries to use stochastic complexity to choose among model classes. The stochastic complexity of the saturated library $\mathcal{B}([0,1]^n)$ is $n$, agreeing with the expected stochastic complexity of the very different, "null" library $\mathcal{B}(\{\frac{1}{2}\})$ which fixes $\theta_i = \frac{1}{2}$ for all $i$. On average, stochastic complexity cannot distinguish the saturated library that varies $\theta$ to match each observation from a dogmatic "null" library that treats the data as i.i.d. noise. Models that treat the data as pure signal have the same stochastic complexity (on average) as those which treat the data as pure noise. Rissanen [2000] encounters such ambiguity between "signal" and "noise" when using MDL in the denoising problem where the dimension of the class of models is on the order of $n$.

## 8.4 Complexity of the Gaussian Library

The parametric complexity of many libraries is unbounded, and as a result one must deviate from the clean definition of stochastic complexity that we have illustrated so far. Perhaps the most important cases of this phenomenon are the Gaussian libraries $\mathcal{G}(\Theta)$ introduced in (8.3). The codebooks in a Gaussian library model $\mathbf{y}$ as though it were a realization of random variables $Y_i \overset{\text{i.i.d.}}{\sim} N(\mu,\sigma^2)$. A Gaussian library cannot be converted into an encyclopedia like those representing a Bernoulli library $\mathcal{B}$. The asymptotic approximation to the parametric complexity (8.7) reveals the problem: the Fisher information (8.8) for $\mu$ is constant but the natural range for this parameter is $\mathbb{R}$. To see the problem more clearly, though, we will avoid this approximation and work directly from the definition.

Assume for the moment that $\sigma^2 = 1$ is known and focus on the one-parameter library with unknown mean,

$$\mathcal{G}(\mathbb{R}) = \{p_\mu : p_\mu(\mathbf{y}) = \frac{e^{-\sum(y_i-\mu)^2/2}}{(2\pi)^{n/2}},\ \mu \in \mathbb{R}\} \ .$$

Following [Barron et al. 1998], the parametric complexity is most easily found by once again using the sufficiency of the sample average $\overline{Y} = \sum Y_i/n$ for $\mu$. Modeled as a sample of normals, the distribution of $\mathbf{y}$ factors into

$$p_\mu(\mathbf{y}) = p_\mu(\overline{y})\ p(\mathbf{y}|\overline{y})$$

where $p(\mathbf{y}|\overline{y})$ is the conditional distribution of $\mathbf{y}$ given $\overline{Y}$, and thus is free of $\mu$. The distribution of the sufficient statistic is $N(\mu, \sigma^2/n)$,

$$p_\mu(\overline{y}) = \left(\frac{n}{2\pi\sigma^2}\right)^{1/2} e^{-\frac{n}{2\sigma^2}(\overline{y}-\mu)^2} \ .$$

When we set $\mu = \overline{y}$ in the NML distribution, this density reduces to a constant,

$$p_{\overline{y}}(\overline{y}) = \left(\frac{n}{2\pi\sigma^2}\right)^{1/2} \ . \tag{8.12}$$

Since the parametric complexity $\log C(\mathcal{G}, \mathbb{R}, \mathbb{R}^n)$ (with both $\Theta$ and $D$ set to the real line) is unbounded, we cannot use stochastic complexity as defined as a criterion for model selection.

One approach to this dilemma is to bound the parametric complexity by constraining $\Theta$. For example, the parametric complexity is finite if we constrain $\Theta$ to the ball of radius $R > 0$ around the origin, $\Theta_R = \{\mu : -R \le \mu \le R\}$. It is important to note that $R$ is a constant chosen prior to looking at the data. This constraint has no effect on the range of data; it only limits the values allowed for $\mu$ and its MLE,

$$\hat{\mu}_R(\mathbf{y}) = \begin{cases} -R, & \overline{y} < -R \ , \\ \overline{y}, & -R \le \overline{y} \le R \ , \\ R, & R < \overline{y} \ . \end{cases}$$

The parametric complexity of $\mathcal{G}(\Theta_R)$ is then 1 plus a multiple of the radius of the parameter space,

$$\begin{aligned} C(\mathcal{G}, \Theta_R, \mathbb{R}^n) &= \int_{\mathbb{R}^n} p_{\hat{\mu}_R(\mathbf{y})}(\overline{y}) p(\mathbf{y}|\overline{y}) d\mathbf{y} \\ &= 2 \int_R^\infty \left(\frac{n}{2\pi\sigma^2}\right)^{1/2} e^{-(\overline{y}-R)^2/2} \, d\overline{y} + \int_{-R}^R \left(\frac{n}{2\pi\sigma^2}\right)^{1/2} \, d\overline{y} \\ &= 1 + \frac{2\sqrt{n}R}{\sqrt{2\pi\sigma^2}} \ . \end{aligned} \tag{8.13}$$

The addition of 1 in (8.13) arises from integrating over those $\mathbf{y}$ for which the MLE lies on the boundary of $\Theta_R$. The associated stochastic complexity for an arbitrary $\mathbf{y} \in \mathbb{R}^n$ is then

$$\begin{aligned} L_{\mathcal{G}(\Theta_R)}(\mathbf{y}) &= \log C(\mathcal{G}, \Theta_R, \mathbb{R}^n) + \log 1/p_{\hat{\mu}_R(\mathbf{y})}(\mathbf{y}) \\ &= \log(1 + \frac{2\sqrt{n}R}{\sqrt{2\pi\sigma^2}}) + \log 1/p_{\overline{y}}(\mathbf{y}) + K(p_{\overline{y}} \| p_{\hat{\mu}_R}) \ . \end{aligned}$$

The last term $K(p_{\overline{y}} \| p_{\hat{\mu}})$ is the Kullback-Leibler divergence between the distribution $p_{\hat{\mu}}$, which uses the constrained MLE, and $p_{\overline{y}}$, which uses the unconstrained sample average,

$$K(p \| q) = \int_{D^n} p(\mathbf{y}) \ \log \frac{p(\mathbf{y})}{q(\mathbf{y})} \ d\mathbf{y} \ .$$

This approach allows us to use stochastic complexity as before, with a single encyclopedia representing a library. The sender and receiver can agree to a particular choice of $R$ prior to encoding $\mathbf{y}$. Stine and Foster [2000] label (8.13) the *unconditional parametric complexity.*

This unconditional approach introduces a problem, however, into the use of stochastic complexity as the criterion for MDL. One must decide prior to observing $\mathbf{y}$ how to constrain $\Theta$. Restricting $\mu$ to lie in $\Theta_R$ may seem natural, but certainly other choices are possible. [Stine and Foster 2000] propose a competitive analysis to pick optimal constraints, but here we consider an alternative method that bounds the parametric complexity in a rather different manner. This alternative *constrains the data* rather than the parameter space.

The most common method for bounding the parametric complexity constrains the data space $D^n$ rather than $\Theta$. Let $D^n_R$ denote the subset of $\mathbb{R}^n$ for which the average of $\mathbf{y}$ lies inside $\Theta_R$,

$$D^n_R = \{\mathbf{y} : \mathbf{y} \in \mathbb{R}^n, -R \leq \overline{y} \leq R\} \ . \tag{8.14}$$

Under this constraint, the normalizing constant becomes

$$C(\mathcal{G}, \mathbb{R}, D^n_R) = \int_{D^n_R} p_{\overline{y}}(\mathbf{y}) \, d\mathbf{y} = \frac{2\sqrt{n}R}{\sqrt{2\pi\sigma^2}} \ , \tag{8.15}$$

which is one less than the constant obtained by constraining the parameter space. Notice that restricting $\mathbf{y}$ to $D^n_R$ implies a constraint on $\Theta$ as well,

$$C(\mathcal{G}, \mathbb{R}, D^n_R) = C(\mathcal{G}, \Theta_R, D^n_R) \ .$$

To distinguish such implicit constraints on $\Theta$ from those set externally, our notation omits the implicit constraints on $\Theta$ when induced by those placed on $\mathbf{y}$.

When constraining $\mathbf{y}$, one must ensure that $\mathbf{y}$ lies in $D^n_R$ or else the library lacks a codebook for the data. Thus, in applications, one replaces the a priori bound $R$ by a data-dependent constraint, say $R(\mathbf{y})$. $R(\mathbf{y})$ is usually chosen so that the unconstrained MLE lies in the implicit parameter space, $\overline{y} \in \Theta_{R(\mathbf{y})}$. This measure of complexity, however, ignores the fact that the receiver needs to know $\overline{y}$. A feature of $\mathbf{y}$ has "leaked out" of the normalization process and must be encoded separately. Constraining $\Theta$ directly produces a "one-volume" encyclopedia that generates a prefix code for $\mathbf{y}$. Constraining the data space $D^n$ leads to a "multi-volume" encyclopedia that cannot generate a prefix code — the receiver does not know which of the volumes to use to decode the message. Consequently, one must add to the stochastic complexity the length of a prefix that identifies $R(\mathbf{y})$,

$$L(\mathbf{y}; \mathcal{G}, \mathbb{R}, D^n_{R(\mathbf{y})}) = \ell(R(\mathbf{y})) + \log\left(\frac{2\sqrt{n}R(\mathbf{y})}{\sqrt{2\pi\sigma^2}}\right) + \log 1/p_{\overline{y}}(\mathbf{y}) \ .$$

The length of the code for $R(\mathbf{y})$ lies outside the framework of the underlying NML model, and thus this approach sacrifices its minimax optimality. In a one-parameter model, the addition of a code for $R(\mathbf{y})$ has little effect on the selection of a model by MDL, especially when formulated along the lines of, say, $R(\mathbf{y}) = 2^{2k(\mathbf{y})}$ for some integer $k(\mathbf{y})$ as in [Rissanen 2000]. The next section shows, however, that the impact of "leaking information" outside the NML normalization grows as one adds more parameters.

Before moving to models of large dimension, the presence of data-dependent bounds introduces other problems as well. In particular, the form of the data-driven constraints can determine whether a library has infinite or finite complexity. We illustrate this aspect of data-driven constraints by introducing an unknown variance $\sigma^2$. To avoid singularities in the likelihood, it is natural to bound $\sigma^2$ away from zero, say $0 < \sigma_0^2 \le \sigma^2$.

With $\sigma^2$ estimated from $\mathbf{y}$, the parametric complexity depends upon how one constrains $\mathbf{y}$. If $\mathbf{y}$ is constrained so that $\overline{y} \in \Theta_R$, the parametric complexity is infinite unless we introduce an upper bound for $\sigma^2$. Barron et al. [1998] and Hansen and Yu [2001] employ this type of constraint. If instead $\mathbf{y}$ is constrained by restricting $\overline{y}$ to a region defined on a standardized scale, say $\overline{y} \in \Theta_{z\sigma/\sqrt{n}}$ as in [Rissanen 1999], then the parametric complexity is finite, *without* the need for an upper bound on $\sigma^2$. This effect of the "shape" of the constraints does not appear if we constrain the parameter space rather than the data.

We begin again with the factorization of the likelihood $p_{\mu,\sigma^2}(\mathbf{y})$ implied by sufficiency. The statistics $\overline{Y}$ and $S^2 = \sum(Y_i - \overline{Y})^2/n$ are independent and jointly sufficient for $\mu$ and $\sigma^2$. The Gaussian likelihood thus factors into a product of three terms,

$$p_{\mu,\sigma^2}(\mathbf{y}) = p(\mathbf{y}|\overline{y}, s^2)\, p_{\mu,\sigma^2}(\overline{y})\, p_{\sigma^2}(s^2) \ ,$$

where $p_{\sigma^2}(s^2)$ denotes the chi-squared density of $S^2$,

$$\begin{aligned} p_{\sigma^2}(s^2) &= \frac{\left(\frac{ns^2}{\sigma^2}\right)^{\alpha-1} e^{-ns^2/2\sigma^2}}{\Gamma(\alpha)2^\alpha}\,\frac{n}{\sigma^2} \\ &= \frac{c_n}{\sigma^2}\left(\frac{s^2}{\sigma^2}\right)^{\alpha-1} e^{-ns^2/2\sigma^2} \ , \end{aligned} \tag{8.16}$$

where the constants $c_n$ and $\alpha$ are

$$c_n = \frac{n^\alpha}{\Gamma(\alpha)2^\alpha} \ , \quad \alpha = \frac{n-1}{2} \ . \tag{8.17}$$

The conditional density of the data $p(Y|\overline{Y}, S^2)$ given $\overline{Y}$ and $S^2$ is free of $\mu$ and $\sigma^2$.

Now let $\hat{D}^n$ denote a subset of $\mathbb{R}^n$ for which the MLE lies within $\hat{\Theta}$. Given this constraint, the parametric complexity is the log of the following integral:

$$\begin{aligned} C(\mathcal{G}, \Theta, \hat{D}^n) &= \int_{\hat{\Theta}} \int_{\hat{D}^n} p(\mathbf{y}|\overline{y}, s^2) p_{\overline{y},s^2}(\overline{y})\, p_{s^2}(s^2) d\mathbf{y}\, d\overline{y}\, ds^2 \\ &= k_n \int_{\hat{\Theta}} \left(\frac{1}{s^2}\right)^{3/2} d\overline{y}\, ds^2 \ , \end{aligned} \tag{8.18}$$

where $k_n$ collects constants from the chi-squared and normal densities,

$$k_n = c_n \, \frac{\sqrt{n} e^{-n/2}}{\sqrt{2\pi}} = \frac{n^{\alpha+1/2} e^{-n/2}}{\sqrt{2\pi}\,\Gamma(\alpha)\,2^{\alpha}} \; . \tag{8.19}$$

To see how the form of the constraints affects the parametric complexity, we just plug them into the integral (8.18) and evaluate. With $\mathbf{y}$ constrained so that $\overline{y} \in \Theta_R$ and $s^2 \geq \sigma_0^2$, the integral splits as

$$\int p_{\overline{y},s^2(\mathbf{y})}(\mathbf{y}) d\mathbf{y} = k_n \int_{-R}^{R} d\overline{y} \int_{\sigma_0^2}^{\infty} \left(\frac{1}{s^2}\right)^{3/2} ds^2 = k_n \frac{2\,R}{\sigma_0^2} \; .$$

The conditional parametric complexity is finite. On the other hand, with $\mathbf{y}$ constrained so that $\overline{y}$ lies within a standardized range (e.g., we plan to encode data whose mean lies within 20 standard errors of zero), the parametric complexity is infinite,

$$\int p_{\overline{y},s^2(\mathbf{y})}(\mathbf{y}) d\mathbf{y} = k_n \int_{\sigma_0^2}^{\infty} \int_{-zs/\sqrt{n}}^{zs/\sqrt{n}} \left(\frac{1}{s^2}\right)^{3/2} d\overline{y}\, ds^2 = 2k_n z \int_{\sigma_0^2}^{\infty} \frac{1}{s^2}\, ds^2 \; .$$

One can bound the complexity in this case by adding a further constraint to the data that restricts $\mathbf{y}$ to those sequences for which, say, $s^2 \leq \sigma_1^2$.

Bounding the parametric complexity by constraining $\mathbf{y}$ thus gives two rather different measures of the complexity of these Gaussian libraries. Consider the effect of restricting $\mathbf{y}$ to those sequences for which $s^2 \leq \sigma_1^2$. If $\mathbf{y}$ is also constrained so that $\overline{y}$ is bounded on the standardized scale, the parametric complexity is a multiple of $\log\left(\sigma_1^2/\sigma_0^2\right)$. If $\overline{y}$ is bounded directly, the parametric complexity is a multiple of $1/\sigma_0^2 - 1/\sigma_1^2$. One tends to infinity with $\sigma_1^2$, whereas the other remains finite.

Unconditional bounds, in contrast, give the same answer whether $\mu$ is restricted directly or on a standardized scale. In either case, the parametric complexity is unbounded. Denote the constrained parameter space by

$$\Theta_R^{\sigma_1^2} = \{(\mu, \sigma^2) : \; -R \leq \mu \leq R, \sigma_0^2 \leq \sigma^2 \leq \sigma_1^2\} \; .$$

Let $\hat{\theta}$ denote the MLE for this space. These constraints are "rectangular" in the sense that

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = \left(\min(\max(-R, \overline{y}), R), \min(\max(\sigma_0^2, s^2), \sigma_1^2)\right) \; .$$

If $(\overline{y}, s^2)$ lies outside of $\Theta_R^{\sigma_1^2}$, then one obtains the MLE by projecting this point perpendicularly onto the boundary of $\Theta_R^{\sigma_1^2}$. When $(\overline{y}, s^2)$ violates both constraints, the projected point is a "corner" of $\Theta_R^{\sigma_1^2}$ [e.g., one corner is $(R, \sigma_1^2)$]. For these rectangular bounds, the normalizing constant is

$$\begin{aligned} C(\mathcal{G}, \Theta_R^{\sigma_1^2}, \mathbb{R}^n) &= \int_{\mathbb{R}^n} p_{\hat{\mu},\hat{\sigma}^2}(\mathbf{y}) \; d\mathbf{y} \\ &\geq \int_{\mathbf{y}:\sigma_0^2 \leq s^2 \leq \sigma_1^2} p_{\hat{\mu},s^2}(\mathbf{y}) \; d\mathbf{y} \end{aligned}$$

$$\begin{aligned}
&= \int_{\sigma_0^2}^{\sigma_1^2} p_{s^2}(s^2)\left(1+\frac{2\sqrt{n}R}{\sqrt{2\pi s^2}}\right) ds^2 \\
&= c_n e^{-n/2}\int_{\sigma_0^2}^{\sigma_1^2} \frac{1}{s^2} ds^2 + 2k_n R\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right) \\
&= c_n e^{-n/2}\left(\log\frac{\sigma_1^2}{\sigma_0^2}\right) + 2k_n R\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right) .
\end{aligned}$$

Notice that $C(\mathcal{G},\Theta_R^{\sigma_1^2},\mathbb{R}^n)$ has both the log of the ratio of the bounds for $\sigma^2$ as well as the difference of the ratios. Thus, $C(\mathcal{G},\Theta_R^{\sigma_1^2},\mathbb{R}^n)$ tends to infinity with $\sigma_1^2$.

A similar calculation shows that the normalizing constant also tends to infinity when the constraints for $\mu$ are specified on the standardized scale. If we restrict $\mu$ to $\Theta_{z\sigma/\sqrt{n}}$, the projections of $(\overline{y},s^2)$ onto the parameter space are no longer rectangular. Nonetheless, we can show that the normalization again tends to infinity. Regardless of the location of $\overline{y}$, the probability at the MLE is at least as large as that at a restricted location, $p_{\hat{\mu},\hat{\sigma}^2} \geq p_{0,\hat{\sigma}^2}$. Consequently, the normalizing constant is bounded below as follows:

$$\begin{aligned}
C(\mathcal{G},\Theta_{z\sigma/\sqrt{n}}^{\sigma_1^2},\mathbb{R}^n) &= \int_{\mathbb{R}^n} p_{\hat{\mu},\hat{\sigma}^2}(\mathbf{y})\ d\mathbf{y} \\
&\geq \int_{\mathbb{R}^n} p_{0,\hat{\sigma}^2}(\mathbf{y})\ d\mathbf{y} \\
&\geq \int_{\mathbf{y}:\sigma_0^2\leq s^2\leq\sigma_1^2} p_{0,s^2}(\mathbf{y})\ d\mathbf{y} \\
&= \int_{\sigma_0^2}^{\sigma_1^2} p_{s^2}(s^2) ds^2 \\
&= c_n e^{-n/2}\ \log\frac{\sigma_1^2}{\sigma_0^2} .
\end{aligned}$$

Again, the normalizing constant tends to infinity as $\sigma_1^2$ grows.

## 8.5 Complexity of Libraries with High Dimension

We consider the use of MDL in the so-called denoising problem. In denoising, the response $\mathbf{y}$ is modeled as a weighted average of selected orthogonal signal vectors $\{W_j^n\}_{j=1}^n$ plus Gaussian noise,

$$\mathbf{Y} = \sum_{j\in\gamma}\beta_j W_j^n + \sigma\epsilon^n\ , \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)\ . \tag{8.20}$$

The range of the summation is a set of indices $\gamma \subset \{1,\ldots,n\}$ that indicates which of the signal vectors have nonzero coefficients. If $j\in\gamma$, then $W_j^n$ affects $\mathbf{y}$; otherwise, the inclusion of $W_j^n$ only adds noise to a fitted reconstruction. The signal vectors might be wavelets, sines and cosines, or any other orthogonal basis for $\mathbb{R}^n$. The problem in denoising is to identify $\gamma$; ideally, the reconstruction requires only a small subset of the signal vectors when the basis is well-chosen. Thus, denoising amounts to variable selection in an orthogonal regression in which one has just as many *possible* predictors as observations.

Because the signal vectors are orthogonal, these models have a convenient canonical form. We can rotate any family of signal vectors into the standard basis for $\mathbb{R}^n$ in which $W_j^n = e_j^n = (0, \ldots, 1_j, 0, \ldots, 0)$. When (8.20) is re-expressed in this way, the underlying probability model becomes the multivariate normal location model, $\mathbf{Y} = \mu^n + \sigma\,\epsilon$ for $\mu^n \in \mathbb{R}^n$. The libraries used in denoising thus generalize the i.i.d library (8.3) by greatly expanding the parameter space

$$\mathcal{G}(\Theta) = \{p_{\mu^n} : p_{\mu^n}(\mathbf{y}) = \frac{e^{-\sum(y_i-\mu_i)^2/2}}{(2\pi)^{n/2}}\ , \mu^n \in \Theta\}\ . \tag{8.21}$$

Each codebook in $\mathcal{G}(\Theta)$ describes $\mathbf{y}$ as a collection of independent, normal random variables, $Y_i \sim N(\mu_i, 1), \quad i = 1, \ldots, n$. The trick is to figure out which $\mu_i \neq 0$. We include the saturated library that allows one parameter per observation and duck some of the boundary problems described in the prior section by fixing $\sigma^2 = 1$. Obviously, one would need to estimate $\sigma^2$ in an application. In wavelet denoising [Donoho and Johnstone 1994], $\sigma^2$ can be estimated quite well from the coefficients of the $n/2$ most localized basis elements.

Our interest lies in using stochastic complexity as the criterion for picking the best dimension for the parameter space. As a first step, consider libraries that are identified by a given set $\gamma$ of nonzero means. Let $\Theta_\gamma = \{\mu^n : \mu^n \in \mathbb{R}^n, \mu_i = 0, i \notin \gamma\}$ denote a $q = |\gamma|$ dimension subspace of $\mathbb{R}^n$. If $\gamma^c$ denotes the complement of $\gamma$ relative to $\{1, 2, \ldots, n\}$, then $\mathcal{G}(\Theta_\gamma)$ contains codebooks for the following models:

$$\mathcal{G}(\Theta_\gamma) = \{p_{\mu^n} : p_{\mu^n}(\mathbf{y}) = \frac{e^{-(\sum_\gamma (y_i-\mu_i)^2 + \sum_{\gamma^c} y_i^2)/2}}{(2\pi)^{n/2}}\} \tag{8.22}$$

Given $\gamma$, we can introduce constraints like those considered in the prior section to obtain the parametric complexity. It remains to identify $\gamma$. If we think of representing $\gamma$ using a vector of Boolean indicators, then the ideas of Section 8.2 become relevant. The stochastic complexity of $\mathcal{B}([0,1])$ for an observed sequence of $n$ i.i.d. Boolean random variables is approximately $\frac{1}{2}\log n + \log\binom{n}{q}$. If we presume $\gamma$, then the resulting stochastic complexity omits the cost of identifying the coordinates of the nonzero parameters.

Rissanen [2000] handles this task by presuming all $2^n$ models are equally likely and adds an $n$-bit code for $\gamma$ to the complexity of $\mathcal{G}(\Theta_\gamma)$ for all $\gamma$. Because this addition adds the same amount to the stochastic complexity for every parameter space, it has no effect on the selection of the best library. This approach does, however, imply a strong bias toward models with about $n/2$ nonzero parameters, as though $\gamma_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\frac{1}{2})$. If instead we incorporate more of $\gamma$ into the NML normalization, we discover that stochastic complexity adapts to the number of nonzero parameters.

One way to retain more of the complexity with the NML normalization is to presume one has an a priori ordering of the basis elements, for example [Barron et al. 1998]. This approach is adopted, for example, when MDL is used to pick the order of a nested sequence of polynomial regressions. Typically, one does not compare all possible polynomials, but rather only compares an increasing sequence of nested models: a linear model to a quadratic model, a quadratic to a cubic, and so forth. For the canonical denoising problem, this knowledge is equivalent to being given an ordering of the parameters, say,

$$\mu_{(1)}^2 \leq \mu_{(2)}^2 \leq \cdots \leq \mu_{(n)}^2 \ .$$

While natural for polynomial regression, such knowledge seems unlikely in denoising.

To retain the coordinate identification within an encyclopedia, we aggregate indexed libraries $\mathcal{G}(\Theta_\gamma)$ into larger collections. Again, let $q = |\gamma|$ denote the number of nonzero parameters and let

$$\Theta_q = \cup_{|\gamma|=q} \Theta_\gamma$$

denote the union of $q$-dimensional subspaces of $\mathbb{R}^n$. Our goal is to select the best of aggregated library $\mathcal{G}(\Theta_q)$. Said differently, our representative encyclopedia has a volume for each $q = 1, \ldots, n$. The use of such an encyclopedia for coding requires only $q$, not $\gamma$ itself, to be specified externally. Because any reasonable code for positive integers assigns roughly equal-length codes to $q = 15$ and $q = 16$, say, the leakage of $q$ outside of the encyclopedia has minimal effect on the use of stochastic complexity in MDL. We can encode $q$ in $O(\log n)$ bits, whereas $\gamma$ requires $O(n)$ bits.

Like other Gaussian libraries, the parametric complexity of $\mathcal{G}(\Theta_q)$ is unbounded without constraints. To specify these, let

$$y_{(1)}^2 < y_{(2)}^2 < \cdots < y_{(n)}^2$$

denote the data ordered in increasing magnitude. The MLE $\hat{\mu}_q^n \in \Theta_q$ matches the largest $q$ elements $y_{(n-q+1)}, \ldots, y_{(n)}$ and sets the others to zero, implying

$$p_{\hat{\mu}_q^n}(\mathbf{y}) = \frac{e^{-(y_{(1)}^2 + \cdots + y_{(n-q)}^2)/2}}{(2\pi)^{n/2}} \ .$$

In order to bound the parametric complexity, we constrain $\mathbf{y}$. For $x \in \mathbb{R}^n$, let $\|x\|^2 = \sum_i x_i^2$ denote the Euclidean norm. Following [Rissanen 2000], we constrain the data to those $\mathbf{y}$ for which the MLE lies in a ball of radius $\sqrt{q}R$ around the origin,

$$D_{q,R}^n = \{\mathbf{y} : \|\hat{\mu}_q^n(\mathbf{y})\| \leq \sqrt{q}\, R\} \ .$$

As with one-dimensional Gaussian models, a prefix code must include a code for $R$ as well as $q$ to identify the appropriate encyclopedia. (For denoising, $q\, R^2$ constrains the "regression sum of squares" that appears in the numerator of the standard $F$-test of a least squares regression. In particular, $R^2$ is *not* the R-squared statistic often seen in regression output.)

The parametric complexity of the library $\mathcal{G}(\Theta_q)$ is the log of the integral of the maximum likelihood density over the restricted range $D^n_{q,R}$. We estimate this complexity by partitioning the normalizing integral into disjoint subsets for which the same coordinates form $\hat{\mu}^n_q$. The subset of nonzero parameters $\gamma$ is fixed over each of these subsets, and the integrals over these subsets are identical. Since there are $\binom{n}{q}$ partitions of the indices that fix $\gamma$, the parametric complexity is $\binom{n}{q}$ times the integral for the convenient subset in which the maximum of the $n-q$ smaller elements, $m_q(\mathbf{y}) = \max(y_1^2, \ldots, y_{n-q}^2)$, is smaller than the minimum of the $q$ larger elements, $M_q(\mathbf{y}) = \min(y_{n-q+1}^2, \ldots, y_n^2)$. Note that $m_q(\mathbf{y}) < M_q(\mathbf{y})$. We then obtain

$$\begin{aligned} C\left(\mathcal{G}, \Theta_q, D^n_{q,R}\right) &= \int_{D^n_{q,R}} p_{\hat{\mu}^n_q}(\mathbf{y})\, d\mathbf{y} \\ &= \binom{n}{q} \int_{\|y_{n-q+1},\ldots,y_n\|^2 < qR^2} \frac{F_{n-q}(M_q(\mathbf{y}))}{(2\pi)^q} dy_{n-q+1} \cdots dy_n \,, \end{aligned} \tag{8.23}$$

where $F_k(x)$ is the integral

$$F_k(x) = \int_{y_1^2,\ldots,y_k^2 < x} \frac{e^{-(y_1^2+\cdots+y_k^2)/2}}{(2\pi)^{k/2}} dy_1 \cdots dy_k \,. \tag{8.24}$$

This integral resembles the cumulative distribution of a chi-squared random variable, but the range of integration is “rectangular” rather than spherical.

The presence of a partition between the largest $q$ elements of $\mathbf{y}$ and the remaining $n-q$ elements in this integration makes it difficult to compute the exact stochastic complexity, but we can still get useful upper and lower bounds. The upper bound is easier to find, so we start there. If we expand the range of integration in $F_{n-q}(x)$ to all of $\mathbb{R}^{n-q}$, the integral is just that of a $q$-dimensional normal density and so $F_{n-q}(x) \le 1$. Thus, for this bound the inner integral expressed as $F_{n-q}$ in (8.23) is just 1, and the constraints together with the binomial coefficient give an upper bound for the normalizing constant,

$$\begin{aligned} C\left(\mathcal{G}, \Theta_q, D^n_{q,R}\right) &\le \binom{n}{q} \int_{\|y_{n-q+1},\ldots,y_n\|^2 < qR^2} dy_{n-q+1} \cdots dy_n \\ &= \binom{n}{q} V_q(\sqrt{q}R) \,, \end{aligned} \tag{8.25}$$

where $V_k(r)$ denotes the volume of the ball of radius $r$ in $\mathbb{R}^k$,

$$V_k(r) = \frac{r^k \pi^{k/2}}{(\frac{k}{2})!} \,.$$

The lower bound results from further constraining the range of integration in (8.23). Rather than integrate over all boundaries between the smaller $n-q$ terms and the larger $q$, we integrate over a single boundary at $2\log(n-q)$, $m_q(\mathbf{y}) \le 2\log(n-q) \le M_q(\mathbf{y})$. The choice of $2\log(n-q)$ as the point of separation follows from the observation that $2\log(n-q)$ is an almost sure bound for the largest squared normal in a sample of $n - q$. If $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} N(0,1)$ and we set

$$P(\max(Z_1^2, \ldots, Z_n^2) < 2 \log n) = \omega_n \ ,$$

then $\lim_{n \to \infty} \omega_n = 1$ [Leadbetter, Lindgren, and Rootzen 1983]. It follows that

$$\begin{aligned} C\left(\mathcal{G}, \Theta_q, D^n_{q,R}\right) &\geq \binom{n}{q} \int_{2\log(n-q) \leq \|y_{n-q+1}, \ldots, y_n\|^2 < q\, R^2} dy_{n-q+1} \cdots dy_n \\ &= \binom{n}{q} A_q(\sqrt{2\log(n-q)}, \sqrt{q} R) \ , \end{aligned} \tag{8.26}$$

where $V_q(r_1, r_2)$ with two arguments denotes the volume of the annulus of inner radius $r_1$ and outer radius $r_2$ in $\mathbb{R}^q$,

$$V_q(r_1, r_2) = V_q(r_2) - V_q(r_1) \ .$$

Combining (8.25) with (8.26), the parametric complexity of the model class with $q$ nonzero parameters is bounded between

$$\binom{n}{q} V_q(\sqrt{2\log n}, \sqrt{q} R) \leq C\left(\mathcal{G}_q, \Theta_q, D^n_{q,R}\right) \leq \binom{n}{q} V_q(\sqrt{q} R) \ . \tag{8.27}$$

A further approximation to these bounds provides insight into the contribution of parameters to the stochastic complexity of high-dimensional models. In practice, a data-driven constraint, say $R(\mathbf{y})$, replaces $R$ to ensure the encyclopedia can encode $\mathbf{y}$. For $q$ of moderate size, the volume of the annulus in the lower bound of (8.27) is small in comparison to that of the ball itself; heuristically, most of the volume of a sphere in $\mathbb{R}^q$ lies near the surface of the sphere. Following this line of reasoning and approximating the logs of factorials as $\log k! \approx k \log k$ (omitting constants unaffected by $q$), we obtain an expression for the parametric complexity that is easy to interpret,

$$\begin{aligned} \log C\left(\mathcal{G}, \Theta_q, D^n_{q,R(\mathbf{y})}\right) &\approx \log \binom{n}{q} + q \log R(\mathbf{y}) \\ &\approx q \log \frac{n}{q} + q \log R(\mathbf{y}) \ , \end{aligned} \tag{8.28}$$

which is reasonable for $q \ll n$.

Consider two situations, one with $q$ large, nonzero $\mu_i$ and the other with $q$ smaller, nonzero parameters. For the "strong-signal" case, assume that the nonzero parameters in $\mu^n$ are all much larger than the almost sure bound $\sqrt{2 \log n}$. In particular, assume that these $\mu_i = O(\sqrt{n})$,

$$\text{Strong signal:} \quad \mu_i^2 \approx c\, n \ , \quad i \in \gamma, \qquad \Rightarrow \quad R^2 = c\, n \ .$$

For the "weak-signal" case, we assume the effects are all near the noise threshold,

$$\text{Weak signal:} \quad \mu_i^2 \approx 2 \log n \ , \quad i \in \gamma, \qquad \Rightarrow \quad R^2 = c \log n \ .$$

For coding data with strong signal, the approximation (8.28) to the parametric complexity resembles the approximation obtained by the standard asymptotic analysis of models with small, fixed dimension. In particular, $\frac{q}{2}\log n$ dominates the approximation (8.28) if $R^2 = O(n)$. This similarity is natural. Given a fixed, parametric model with finitely many parameters, the standard analysis lets $n \to \infty$ while holding the model fixed. Thus, the estimation problem becomes much like our strong-signal case: with increasing samples, the standard errors of estimates of the fixed set of parameters fall at the rate of $1/\sqrt{n}$, and the underlying "true model" becomes evident. The term $q \log R$ dominates the approximation (8.28), implying a penalty of $\frac{1}{2}\log n$ as the model grows from dimension $q$ to $q+1$, just as in (8.7).

The penalty for adding a parameter is rather different when faced with weak signals. In such cases, the approximation (8.28) suggests a penalty that resembles those obtained from adaptive thresholding and empirical Bayes. With $R = O(\log n)$, $q \log n/q$ dominates the approximate parametric complexity (8.28). This type of penalty appears in various forms of so-called adaptive model selection. For choosing $q$ out of $p$ possible parameters, one can motivate an adaptive model selection criterion that contains a penalty of the form $q \log p/q$ from information theory [Foster and Stine 1996], multiple comparisons [Abramovich, Benjamini, Donoho, and Johnstone 2000], and empirical Bayes [George and Foster 2000].

## 8.6 Discussion

So, what is the asymptotic contribution of parameters to the stochastic complexity of a model? Unfortunately, the answer appears to be that "it depends." Ideally, the parametric complexity is a fixed measure of the "complexity" of a class of models, or library. Because the idealized parametric complexity is invariant of $\mathbf{y}$, it offers a clear assessment of how the fit (namely, the maximum of the log-likelihood) of a model can overstate the ability of such models to represent data. In models with rich parameterizations, the parametric complexity sometimes increases at the familiar rate of $\frac{1}{2}\log n$ per parameter (one-parameter Bernoulli, high-signal denoising), but at other times grows dramatically slower. The cost per parameter is only 1 in the saturated Bernoulli model and about $\log n/q$ in low-signal denoising. The latter problem, finding the subtle, yet useful parameters from a large collection of possible effects seems, to us, most interesting and worthy of further study.

Adaptive criteria that vary the penalty for adding a parameter have demonstrated success in applications. For example, we have built predictive models for credit risk that consider on the order of 100,000 features as predictors [Foster and Stine 2002]. The model was identified using a variation on the adaptive rule suggested in the weak-signal denoising problem. Such applications of adaptive rules require other important considerations that we have not addressed here. In particular, modeling with an adaptive rule requires careful estimation of the standard error of parameters. In modeling credit risk, the introduction of several spurious predictors leads to bias in the estimate of the effects of subsequent predictors and a cascade of overfitting.

## References

Abramovich, F., Y. Benjamini, D. Donoho, and I. Johnstone (2000). Adapting to unknown sparsity by controlling the false discovery rate. Technical report 2000–19, Department of Statistics, Stanford University, Stanford, CA.

Barron, A.R., J. Rissanen, and B. Yu (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, *44*, 2743–2760.

Donoho, D.L. and I.M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, *81*, 425–455.

Foster, D.P. and R.A. Stine (1996). Variable selection via information theory. Technical report discussion paper 1180, Center for Mathematical Studies in Economics and Management Science, Northwestern University, Chicago.

Foster, D.P. and R.A. Stine (1999). Local asymptotic coding. *IEEE Transactions on Information Theory*, *45*, 1289–1293.

Foster, D.P. and R.A. Stine (2002). Variable selection in data mining: Building a predictive model for bankruptcy. Submitted for publication.

George, E.I. and D.P. Foster (2000). Calibration and empirical Bayes variable selection. *Biometrika*, *87*, 731–747.

Hansen, M.H. and B. Yu (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, *96*, 746–774.

Leadbetter, M.R., G. Lindgren, and H. Rootzen (1983). *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer-Verlag.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, *11*, 416–431.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, *14*, 1080–1100.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*, 40–47.

Rissanen, J. (1999). Hypothesis selection and testing by the MDL principle. *Computer Journal*, *42*, 260–269.

Rissanen, J. (2000). MDL denoising. *IEEE Transactions on Information Theory*, *46*, 2537–2543.

Shtarkov, Y.M. (1987). Universal sequential coding of single messages. *Problems of Information Transmission 23*, 3–17.

Stine, R.A. and D.P. Foster (2000). The competitive complexity ratio. In *2000 Conference on Information Sciences and Systems*, pp. WP8 1–6. Princeton

University, Princeton, NJ.