

A Minimum Description Length Approach to Multitask Feature Selection

Brian Tomasik

May 2009

Abstract

One of the central problems in statistics and machine learning is *regression*: Given values of input variables, called *features*, develop a model for an output variable, called a *response* or *task*. In many settings, there are potentially thousands of possible features, so that *feature selection* is required to reduce the number of predictors used in the model. Feature selection can be interpreted in two broad ways. First, it can be viewed as a means of *reducing prediction error* on unseen test data by improving model generalization. This is largely the focus within the machine-learning community, where the primary goal is to train a highly accurate system. The second approach to feature selection, often of more interest to scientists, is as a form of *hypothesis testing*: Assuming a “true” model that generates the data from a small number of features, determine which features actually belong in the model. Here the metrics of interest are precision and recall, more than test-set error.

Many regression problems involve not one but several response variables. Often the responses are suspected to share a common underlying structure, in which case it may be advantageous to share information across the responses; this is known as *multitask learning*. As a special case, we can use multiple responses to better identify shared predictive features—a project we might call *multitask feature selection*.

This thesis is organized as follows. Section 1 introduces feature selection for regression, focusing on ℓ_0 regularization methods and their interpretation within a Minimum Description Length (MDL) framework. Section 2 proposes a novel extension of MDL feature selection to the multitask setting. The approach, called the “Multiple Inclusion Criterion” (MIC), is designed to borrow information across regression tasks by more easily selecting features that are associated with multiple responses. We show in experiments on synthetic and real biological data sets that MIC can reduce prediction error in settings where features are at least partially shared across responses. Section 3 surveys hypothesis testing by regression with a single response, focusing on the parallel between the standard Bonferroni correction and an MDL approach. Mirroring the ideas in Section 2, Section 4 proposes a novel MIC approach to hypothesis testing with multiple responses and shows that on synthetic data with significant sharing of features across responses, MIC outperforms standard FDR-controlling methods in terms of finding true positives for a given level of false positives. Section 5 concludes.

1 Feature Selection with a Single Response

The standard *linear regression model* assumes that a response y is generated as a linear combination of m predictor variables (“features”) x_1, \dots, x_m with some random noise:

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \quad (1)$$

where we assume the first feature x_1 is an intercept whose value is always 1. Given n observations of the features and responses, we can write the y values in an $n \times 1$ vector Y and the x values in an $n \times m$ matrix X . Assuming the observations are independent and identically distributed, (1) can be rewritten as

$$Y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}_n(0, \sigma^2 I_{n \times n}), \quad (2)$$

where $\beta = \begin{bmatrix} \beta_1 \\ \dots \\ \beta_m \end{bmatrix}$, \mathcal{N}_n denotes the n -dimensional Gaussian distribution, and $I_{n \times n}$ is the $n \times n$ identity matrix.

The maximum-likelihood estimate $\hat{\beta}$ for β under this model can be shown to be the one minimizing the residual sum of squares:

$$\text{RSS} := (Y - X\beta)'(Y - X\beta). \quad (3)$$

The solution is given by

$$\hat{\beta} = (X'X)^{-1} X'Y \quad (4)$$

and is called the *ordinary least squares* (OLS) estimate for β .

In some cases, we may want to restrict the number of features in our model to a subset of q of them, including the intercept. In this case, we pretend that our X matrix contains only the relevant q columns when applying (4); the remaining $m - q$ entries of the $\hat{\beta}$ matrix are set to 0. I'll denote the resulting estimate by $\hat{\beta}_q$, and the RSS for that model by

$$\text{RSS}_q := (Y - X\hat{\beta}_q)'(Y - X\hat{\beta}_q). \tag{5}$$

1.1 Penalized Regression

In many cases, regression problems have large numbers of potential features. For instance, [FS04] predicted credit-card bankruptcy using a model with more than $m = 67,000$ potential features. In bioinformatics applications, it is common to have thousands or tens of thousands of features for, say, the type of each of a number of genetic markers or the expression levels of each of a number of gene transcripts. The number of observations, in contrast, is typically a few hundred at best.

The OLS estimate (4) breaks down when $m > n$, since the $m \times m$ matrix $X'X$ is not invertible due to having rank at most n . Moreover, it's implausible that a given response is *actually* linearly related to such a large number of features; a model $\hat{\beta}$ with lots of nonzeros is probably overfitting the training data.

The statistics and machine-learning communities have developed a number of approaches for addressing this problem. One of the most common is *regularized regression*, which aims to minimize not (3) directly, but a penalized version of the residual sum of squares:

$$(Y - X\beta)'(Y - X\beta) + \lambda \|\beta\|_p, \tag{6}$$

where $\|\beta\|_p$ represents the ℓ_p norm of β and λ is a tunable hyperparameter, to be determined by cross-validation or more sophisticated *regularization path* approaches.¹

Ridge regression takes the penalty as proportional to the (square of the) ℓ_2 norm: $\lambda \|\beta\|_2^2$. This corresponds to a Bayesian maximum a posteriori estimate for β under a Gaussian prior $\beta \sim \mathcal{N}(0_{m \times 1}, \frac{\sigma^2}{\lambda} I_{m \times m})$, with σ^2 as in (1) [Bis06, p. 153]. Under this formulation, we have

$$\hat{\beta} = (X'X + \lambda I_{m \times m})^{-1} X'Y,$$

which is computationally valid because $X'X + \lambda I_{m \times m}$ is invertible for $\lambda > 0$. However, because the square of a decimal number less than one is much smaller than the original number, the ℓ_2 norm offers little incentive to drive entries of $\hat{\beta}$ to 0—many of them just become very small.

Another option is to penalize by the ℓ_1 norm, which is known as *lasso regression* and is equivalent to a double-exponential prior on β [Tib96]. Unlike ℓ_2 , ℓ_1 regularization doesn't square the coefficients, and hence the entries of $\hat{\beta}$ tend to be sparse. In this way, ℓ_1 regression can be seen as a form of feature selection—i.e., choosing a subset of the original features to keep in the model (e.g., [BL97]). Sparsity helps to avoid overfitting to the training set; as a result, the number of training examples required for successful learning with ℓ_1 regularization grows only logarithmically with the number of irrelevant features, whereas this number grows linearly for ℓ_2 regression [Ng04]. Sparse models also have the benefit of being more interpretable, which is important for scientists who want to know which particular variables are actually relevant for a given response. Building regression models for interpretation is further discussed in Sections 3 and 4.

If ℓ_2 regression fails to achieve sparsity because coefficients are squared, then, say, $\ell_{\frac{1}{2}}$ regression should achieve even more sparsity than ℓ_1 . As p approaches 0, $\|\beta\|_p$ approaches the number of nonzero values in β . Hence, regularization with what is called the " ℓ_0 norm" is *subset selection*: Choosing a small number of the original features to retain in the regression model. Once a coefficient is in the model, there's no incentive to drive it to a small value; all that counts is the cost of adding it in the first place. The ℓ_0 norm has a number of advantages [LPFU08], including bounded worst-case risk with respect to the ℓ_1 norm and better control of a measure called the "false discover rate" (FDR), explained more fully in Section 3. Moreover, as [OTJ09, p. 1] note, "A virtue of the [ℓ_0] approach is that it focuses on the qualitative decision as to whether a covariate is relevant to the problem at hand, a decision which is conceptually distinct from parameter estimation." However, they add, "A virtue of the [ℓ_1] approach is its computational tractability." Indeed, exact ℓ_0 regularization requires subset search, which has been proved NP-hard [Nat95]. In practice, therefore, an approximate greedy algorithm like forward stepwise selection is necessary [BL05, p. 582].

In a regression model, residual sum of squares is proportional up to an additive constant to the negative log-likelihood of β . Therefore, ℓ_0 regularization can be rephrased as a *penalized likelihood* criterion (with a different λ than in (6)):

$$-2 \ln P(Y | \hat{\beta}_q) + \lambda q, \tag{7}$$

¹See, e.g., [Fri08] for an excellent introduction.

where q is the number of features in the model, and $P(Y | \hat{\beta}_q)$ is the likelihood of the data given a model containing q features. As noted in [GF00], statisticians have proposed a number of choices for λ , including

- $\lambda = 2$, corresponding to Mallows' C_p [Mal73], or approximately the Akaike Information Criterion (AIC) [Aka73],
- $\lambda = \ln n$, the Bayesian Information Criterion (BIC) [Sch78], and
- $\lambda = 2 \ln m$, the Risk Inflation Criterion (RIC) [DJ94, FG94].

It turns out that each of these criteria can be derived from an information-theoretic principle called Minimum Description Length (MDL) under different “model coding” schemes (e.g., [FS99, HY99]). MDL forms the focus of this thesis, and its approach to regression is the subject of the next subsection.

1.2 Regression by Minimum Description Length

MDL [Ris78, Ris99] is a method of model selection that treats the best model as the one which maximally compresses a digital representation of the observed data. We can imagine a “Sender” who wants to transmit some data in an email to a “Receiver” using as few bits as possible [Sti04]. In the case of linear regression, we assume that both Sender and Receiver know the $n \times m$ matrix X , and Sender wants to convey the values in the $n \times 1$ matrix Y . One way to do this would be to

send the raw values for each of the n observations $Y = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}$. However, if the response is correlated with some features,

it may be more efficient first to describe a regression model $\hat{\beta}$ for Y given X and then to enumerate the residuals $Y - X\hat{\beta}$, which—having a narrower distribution—require fewer bits to encode.

To minimize description length, then, Sender should choose $\hat{\beta}$ to minimize

$$\mathcal{D}(Y | \hat{\beta}) + \mathcal{D}(\hat{\beta}), \tag{8}$$

where the first term is the description length of the residuals about the model, and the second term is the description length of the model itself.² Exactly what these terms mean will be elaborated below.

1.2.1 Coding the Data: $\mathcal{D}(Y | \hat{\beta})$

The Kraft inequality in information theory [CT06, sec. 5.2] implies that for any probability distribution $\{p_i\}$ over a finite or countable set, there exists a corresponding code with codeword lengths $[-\lg p_i]$. Moreover, these code lengths are optimal in the sense of minimizing the expected code length with respect to $\{p_i\}$. If Sender and Receiver agree on a model for the data, e.g., (1), then they have a probability distribution over possible configurations of residuals ϵ , so they will agree to use a code for the residuals with lengths

$$\mathcal{D}(Y | \hat{\beta}) = -\lg P(\epsilon | \hat{\beta}) = -\lg P(Y | \hat{\beta}), \tag{9}$$

that is, the negative log-likelihood of the data given the model. This is a standard statistical measure for poorness of fit.³

Consider a stepwise-regression scenario in which our model currently contains $q - 1$ features (including the intercept term), and we’re deciding whether to include an extra q^{th} feature. Let Y_i denote the i^{th} entry (row) of Y and $\hat{\beta}_q$ a model with all q features. (2) and (9) give

$$\begin{aligned} \mathcal{D}(Y | \hat{\beta}_q) &= -\lg \prod_{i=1}^n P(Y_i | \hat{\beta}_q) \\ &= -\sum_{i=1}^n \lg \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2\sigma^2} (Y_i - X\hat{\beta}_q)^2 \right) \right] \\ &= \frac{1}{2 \ln 2} \left[n \ln(2\pi\sigma^2) + \frac{\text{RSS}_q}{\sigma^2} \right], \end{aligned} \tag{10}$$

²This is what [Gru05, p. 11] calls the “crude, two-part version of MDL.” Beginning with [Ris86], Rissanen introduced a “one-part” version of MDL based on a concept called “stochastic complexity.” However, it too divides the description length into terms for model fit and model complexity, so it is in practice similar to two-part MDL [Sti04].

³We use “idealized” code lengths and so drop the ceiling on $-\lg P(Y | \hat{\beta})$ [BY98, p. 2746]. Also, since $P(Y | \hat{\beta})$ is a continuous density, we would in practice need to discretize it. This could be done to a given precision with only a constant impact on code length [CT06, sec. 8.3].

where RSS_q is as in (5). Of course, σ^2 is in practice unknown, so we estimate it by⁴

$$\hat{\sigma}^2 = \frac{\text{RSS}_q}{n}. \quad (11)$$

Inserting this estimate for σ^2 in (10) gives

$$\begin{aligned} \mathcal{D}(Y | \hat{\beta}_q) &= \frac{1}{2 \ln 2} [n \ln(2\pi\hat{\sigma}^2) + n] \\ &= \frac{n}{2 \ln 2} \left[\ln \left(\frac{2\pi \text{RSS}_q}{n} \right) + 1 \right]. \end{aligned} \quad (12)$$

Unfortunately, in practice, (11) seems to cause overfitting, especially when used for multiple responses as in Section 2.2.2. The overfitting comes from the fact that $\hat{\sigma}^2$ assumes that the current q^{th} feature actually comes into the model; if this is not the case, the estimated variance is spuriously too low, which allows random noise to appear more unlikely than it really is. To prevent overfitting, we use instead an estimate of $\hat{\sigma}^2$ based on the model without the current feature: $\hat{\beta}_{q-1}$. That is,

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - X \hat{\beta}_{q-1})^2 = \frac{\text{RSS}_{q-1}}{n}, \quad (13)$$

which means that (10) becomes instead

$$\mathcal{D}(Y | \hat{\beta}_q) = \frac{n}{2 \ln 2} \left[\ln \left(\frac{2\pi \text{RSS}_{q-1}}{n} \right) + \frac{\text{RSS}_q}{\text{RSS}_{q-1}} \right]. \quad (14)$$

1.2.2 Coding the Model: $\mathcal{D}(\hat{\beta})$

Just as $\mathcal{D}(Y | \hat{\beta})$ depends on the model for the residuals that Sender and Receiver choose, so their coding scheme for $\hat{\beta}$ itself will reflect their prior expectations.⁵ When the number of features m is large (say, 1000), Sender will likely only want to transmit a few of them that are most relevant, and hence the $\hat{\beta}$ matrix will contain mostly zeros.⁶ So the first step in coding $\hat{\beta}$ could be to say where the nonzero entries are located; if only a few features enter the model, this can be done relatively efficiently by listing the indices of the features in the set $\{1, 2, \dots, m\}$. This requires $\lceil \lg m \rceil$ bits, which we idealize as just $\lg m$.

The second step is to encode the numerical values of those coefficients. Rissanen [Ris83] suggested the basic approach for doing this: Create a discrete grid over some possible parameter values, and use a code for integers to specify which grid element is closest.⁷ [Sti04] described a simple way to approximate the value of a particular coefficient $\hat{\beta}$: Encode an integer version of its z-score relative to the null-hypothesis value β_0 (which in our case is 0):

$$\left\langle \frac{\hat{\beta} - \beta_0}{\text{SE}(\hat{\beta})} \right\rangle = \left\langle \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} \right\rangle, \quad (15)$$

where $\langle x \rangle$ means “the closest integer to x .” The z-score can then be coded with the idealized universal code for the positive integers of [Eli75] and [Ris83], in which the cost to code $i \in \{1, 2, 3, \dots\}$ is

$$\lg^* i + b,$$

⁴ This is the maximum-likelihood estimate for σ^2 , which Sender uses because, ignoring model-coding cost, maximizing likelihood is equivalent to minimizing description length. In practice, of course, many statisticians use the unbiased estimate

$$\hat{\sigma}^2 = \frac{\text{RSS}_q}{n - q}.$$

[Sti04, Appendix A] makes an interesting case for this latter value within the MDL framework. However, we use the maximum-likelihood estimate for its elegance and convenience.

⁵ Again by the Kraft inequality, we can interpret $2^{-\mathcal{D}(\beta)}$ as a prior over possible models β . In fact, this is done explicitly in the Minimum Message Length (MML) principle (e.g., [Wal05]), a Bayesian sister to MDL which chooses the model $\hat{\beta}$ with maximum $P(\beta | Y)$, i.e., the model that minimizes

$$-\lg P(\beta | Y) = -\lg P(\beta) - \lg P(Y | \beta) + \text{const.}$$

⁶ For convenience, we assume that Sender always transmits the intercept coefficient $\hat{\beta}_1$ and that, in fact, doing so is free of charge. Thus, even if Sender were to encode all other feature coefficients as 0, Receiver would at least have the average \bar{Y} of the Y values to use in reconstructing Y .

⁷ Thus, Sender transmits only rounded coefficient values. Rounding adds to the residual coding cost because Sender is not specifying the exact MLE, but the increase tends to be less than a bit [Ris89].

where $\lg^* i := \lg i + \lg \lg i + \lg \lg \lg i + \dots$ so long as the terms remain positive, and $b \approx \lg 2.865 \approx 1.516$ [Ris83, p. 424] is the constant such that

$$\sum_{i=1}^{\infty} 2^{-(\lg^* i + b)} = 1.$$

We require the \lg^* instead of a simple \lg because the number of bits Sender uses to convey the integer i will vary, and she needs to tell Receiver how many bits to expect. This number of bits is itself an integer that can be coded, hence the iteration of logarithms. The middle row of Table 1 shows example costs with this code.

i	1	2	3	4	5	10	100
Universal-code cost for i	1.5	2.5	3.8	4.5	5.3	7.4	12.9
Universal-style cost for i , truncated at 1000	1.2	2.2	3.4	4.2	5.0	7.0	12.6

Table 1: Example costs of the universal code for the integers.

In fact, in practice it’s unnecessary to allow our integer code to extend to arbitrarily large integers. We’re interested in features near the limit of detectability, and we expect our z -scores to be roughly in the range ~ 2 to ~ 4 , since if they were much higher, the true features would be obvious and wouldn’t require sensitive feature selection. We could thus impose some maximum possible z -score Z that we might ever want to encode (say, 1000) and assume that all of our z -scores will fall below it. In this case, the constant c can be reduced to a new value c_Z , now only being large enough that

$$\sum_{i=1}^Z 2^{-(\lg^* i + c_Z)} = 1. \tag{16}$$

In particular, $c_{1000} \approx 1.199$, which is reflected in the reduced costs shown in the third row of Table 1. As it turns out, $c_Z \approx 1$ over a broad range (say, for $Z \in \{5, \dots, 1000\}$).

In our implementation, we avoid computing the actual values of our z -scores (though this could be done in principle), instead assuming a constant cost of 2 bits per coefficient. Though MDL theoretically has no tunable parameters once Sender and Receiver have decided on their models, the number of bits to specify a coefficient can act as one, having much the same effect as the significance level α of a hypothesis test (see Section 3.3 for details). We found 2 bits per coefficient to work well in practice.

1.2.3 Summary

Combining the cost of the residuals with the cost of the model gives the following formula for the description length as a function of the number of features q that we include in the model:

$$-\lg P(Y | \hat{\beta}_q) + q(\lg m + 2). \tag{17}$$

Note the similarity between (17) and the RIC penalty for (7).

2 MIC for Prediction

Standard multivariate linear regression as described in Section 1 assumes many features but only a single response. However, a number of practical problems involve both multiple features and multiple responses. For instance, a biologist may have transcript abundances for thousands of genes that she would like to predict using genetic markers [KCY+06], or she may have multiple diagnostic categories of cancers whose presence or absence she would like to predict using transcript abundances [KWR+01]. [BF97] describe a situation in which they were asked to use over 100 econometric variables to predict stock prices in 60 different industries.

More generally, we suppose we have h response variables y_1, \dots, y_h to be predicted by the same set of features x_2, \dots, x_p (with x_1 being an intercept, as before). In parallel to (2), we can write

$$Y = X\beta + \epsilon, \tag{18}$$

where now Y is an $n \times h$ matrix with each response along a column. X remains $n \times m$, while β is $m \times h$ and ϵ is $n \times h$.

In general, the noise on the responses may be correlated; for instance, if our responses consist of temperature measurements at various locations, taken with the same thermometer, then if our instrument drifted too high at one location, it will have been too high at the other. A second, practical reason we might make this assumption comes from our use of stepwise regression to choose our model: In the early iterations of a stepwise algorithm, the effects of features not present in the model show up as part of the “noise” error term, and if two responses share a feature not yet in the model, the portion of the error term due to that feature will be the same. Thus, we may decide to take the rows of ϵ have nonzero covariance:

$$\epsilon_i \sim \mathcal{N}_h(0, \Sigma), \tag{19}$$

where ϵ_i is the i th row of ϵ , and Σ is an arbitrary (not necessarily diagonal) $h \times h$ covariance matrix. In practice, however, we end up using a diagonal estimate of Σ , as discussed in Section 2.2.2.

2.1 Multitask Learning

The naïve approach to regression with the above model is to treat each response as a separate regression problem and calculate each column of β using just the corresponding column of Y . This is completely valid mathematically; it even works in the case where the covariance matrix Σ is non-diagonal, because the maximum-likelihood solution for the mean of a multivariate Gaussian is independent of the covariance [Bis06, p. 147]. However, intuitively, if the regression problems are related, we ought to be able to do better by “borrowing strength” across the multiple responses. This concept goes by various names; in the machine-learning community, it’s often known as “multitask learning” (e.g., [Car97]) or “transfer learning” (e.g., [RNK06]).

Several familiar machine-learning algorithms naturally do multitask learning; for instance, neural networks can share hidden-layer weights with multiple responses. In addition, a number of explicit multitask regression methods have been proposed, such as *curds and whey*, which takes advantage of correlation among responses [BF97], and various dimensionality-reduction algorithms (e.g., [AM05]).

Some techniques even do explicit feature selection. For instance, [BVF98, BVF02] present Bayesian Markov Chain Monte Carlo approaches for large numbers of features. [ST05] present a *Multiresponse Sparse Regression* (MRSR) algorithm that selects features in a stepwise manner by choosing at each iteration the feature most correlated with the residuals of the responses from the previous model.⁸ We have not explored the above algorithms in detail, but comparing them against MIC could be a fruitful direction for future work. Below we elaborate on two multitask-learning approaches that we have used.

2.1.1 AndoZhang

Ando and Zhang [AZ05] aim to use multiple prediction problems to learn an underlying shared structural parameter on the input space. Their presentation is general, but they give the particular example of a linear model, which we’ll rewrite using the notation of this thesis. Given h tasks and m features per task, the authors assume the existence of an $H \times m$ matrix⁹ Θ , with $H < m$, such that for each task $k = 1, \dots, h$,

$$Y^k = X\beta^k + X\Theta'\beta_\Theta^k,$$

where superscript k denotes the k^{th} column and β_Θ is an $H \times h$ matrix of weights for the “lower dimensional features” created by the product $X\Theta'$ (p. 1824). The transformation Θ is common across tasks and is thus the way we share information. The authors develop an algorithm, referred to as *AndoZhang* in this thesis, that chooses (p. 1825)

$$\left[\left\{ \hat{\beta}, \hat{\beta}_\Theta \right\}, \hat{\Theta} \right] = \underset{\{\beta, \beta_\Theta, \Theta\}}{\operatorname{argmin}} \sum_{k=1}^h \left(\frac{1}{n} \|Y^k - X(\beta^k + \Theta'\beta_\Theta^k)\|^2 + \lambda_k \|\beta^k\|^2 \right) \quad \text{such that } \Theta\Theta' = I_{H \times H}.$$

Here $\{\lambda_k\}_{k=1}^h$ are regularization weights, and (unlike Ando and Zhang) we’ve assumed a constant number n of observations for each task, as well as a quadratic loss function. $\|\cdot\|^2$ denotes the square of the ℓ_2 norm.

Ando and Zhang present their algorithm as a tool for semi-supervised learning: Given one particular problem to be solved, it’s possible to generate auxiliary problems from the known data, solve them, and then use the shared structural parameter Θ on the original problem. Of course, the approach works just as well when we start out with all of the tasks as problems to be solved in their own right.

⁸Another sparse-regression algorithm is BBLasso, described in Section 2.1.2 below.

⁹The authors use h instead of H , but we have already used that variable to denote the number of tasks.

2.1.2 BBLasso

A few multitask learning methods actually do feature selection, i.e., building models that contain only a subset of the original features. AndoZhang, with its ℓ_2 regularization penalty, leaves in most or all of the features, but ℓ_1 methods, such as *BBLasso* presented below, do not.

[AEP08] present an algorithm for learning *functions* of input features across tasks. This is more general than the approach considered in this thesis of choosing merely a subset of the original features, but [AEP08, sec. 2.2] note that their approach reduces to feature selection with an identity function. Regularizing with an ℓ_1 norm over features, their problem thus becomes

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\operatorname{RSS}_{\beta} + \lambda \sum_{i=1}^m \|\beta_i\| \right), \quad (20)$$

where $\operatorname{RSS}_{\beta} = \operatorname{trace}[(Y - X\beta)'(Y - X\beta)]$ is the residual sum of squares for all h responses, and β_i denotes the i^{th} row of β . The i^{th} term in the sum represents the magnitude of coefficients assigned to feature i ; the sum over these values for each i amounts to an ℓ_1 norm over the rows.

The ℓ_2 norm favors shared coefficients across responses for the same features. To see this, [AEP08, p. 6] suggest the case in which the entries of β can only be 0 or 1. If h different features each have a single nonzero response coefficient, the penalty will be $\lambda \sum_{j=1}^h 1 = \lambda h$, since $\|\beta_i\| = 1$ for each such feature i . However, if a single feature shares a coefficient of 1 across all h responses, the penalty is only $\lambda \sqrt{\sum_{j=1}^h 1^2} = \lambda \sqrt{h}$. The same number and magnitude of nonzero coefficients thus leads to a much smaller penalty when the coefficients are shared.

In principle the penalty term can be made even smaller by including nonzero coefficients for only some of the responses. For instance, in the example above, the penalty would be $\sqrt{h}/2$ if just half of the responses had nonzero coefficients (i.e., coefficients of 1). However, in reality, the coefficient values are not restricted to 0 and 1, meaning that unhelpful coefficients, rather than being set to 0, tend to get very small values whose square is negligible in the ℓ_2 -norm penalty. In practice, then, this algorithm adds entire rows of nonzero coefficients to $\hat{\beta}$.

Of course, as [OTJ09, p. 4] note, the ℓ_2 norm used here could be generalized to an ℓ_p norm for any $p \in [1, \infty)$. $p = 1$ corresponds to “no sharing” across the responses, since (20) would then reduce to the objective function for h independent ℓ_1 -penalized regressions. At the opposite extreme, $p = \infty$ corresponds to “maximal sharing”; in that case, only the maximum absolute coefficient value across a row matters.

A number of efficient optimization approaches have been proposed for (20) with $p = 2$ and $p = \infty$; see, e.g., the related-work sections of [TVW05] and [ST07] for a survey. In particular, [OTJ09] propose a fast approximate algorithm using a regularization path, that is, a method which implicitly evaluates (20) at all values of λ .

2.2 Regression by MIC

Following Section 1.2, this section considers ways we might approach regression with multiple responses from the perspective of information theory. One strategy, of course, would be to apply the penalized-likelihood criterion (17) to each response in isolation; we’ll call this the *RIC* method, because the dominant penalty in (17) is $\lg m$, the same value prescribed by the Risk Inflation Criterion for equation (7).

However, following the intuition of BBLasso and other multitask algorithms, we suspect that features predictive of one task are more likely predictive of other, related tasks. For example, if two of our responses are “level of HDL cholesterol” and “level of LDL cholesterol,” we expect that lifestyle variables correlated with one will tend to be correlated with the other. The following multitask coding scheme, called the *Multiple Inclusion Criterion* (MIC), is designed to take advantage of such situations by making it easier to add to our models features that are shared across responses.

2.2.1 Coding the Model

The way MIC borrows strength across responses is to efficiently specify the feature-response pairs in the $m \times h$ matrix $\hat{\beta}$. The naïve approach would be to put each of the mh coefficients in a linear order and specify the index of the desired coefficient using $\lg(mh)$ bits. But we can do better. If we expect nearly all the responses to be correlated with the predictive features, we could give all of the responses nonzero coefficients (using $2h$ bits to code the values for each of the h response coefficients) and simply specify the feature that we’re talking about using $\lg m$ bits, as in Section 1.2.2. We’ll call this the *fully dependent MIC* coding scheme (or *Full MIC* for short). It amounts to adding entire rows of nonzero coefficients at a time to our $\hat{\beta}$ matrix, in much the same way that BBLasso does.

In many cases, however, the assumption that each feature is correlated with almost all of the responses is unrealistic. A more flexible coding scheme would allow us to specify only a subset of the responses to which we want to give nonzero coefficients. For instance, suppose we’re considering feature number 703; of the $h = 20$ responses, we think that only responses

$\{2, 5, 11, 19\}$ should have nonzero coefficients with the current feature. We can use $\lg m$ bits to specify our feature (number 703) once, and then we can list the particular responses that have nonzero coefficients with feature 703, thereby avoiding four different $\lg(mh)$ penalties to specify each coefficient in isolation.

A standard way to code a subset of a set of size h is to first specify how many $k \leq h$ elements the subset contains and then which of the $\binom{h}{k}$ possible subsets with k elements we're referring to [CT06, sec. 7.2]. In particular, we choose to code k using

$$\lg^* k + c_h \tag{21}$$

bits, with c_h as defined in (16).¹⁰ We then need $\lg \binom{h}{k}$ additional bits to specify the particular subset. We refer to this code as *partially dependent MIC* (or *Partial MIC* for short).

2.2.2 Coding the Data

As usual, $\mathcal{D}(Y | \hat{\beta}_q) = -\lg P(Y | \hat{\beta}_q)$ for the model that Sender and Receiver choose, which in this case is (18). If we allow rows of ϵ to have arbitrary covariance as in (19), then

$$\mathcal{D}(Y | \hat{\beta}_q) = \frac{1}{2 \ln 2} \left[n \ln ((2\pi)^h |\Sigma|) + \sum_{i=1}^n (Y_i - X_i \hat{\beta}_q)' \Sigma^{-1} (Y_i - X_i \hat{\beta}_q) \right], \tag{22}$$

with subscript i denoting the i^{th} row. Since Σ is in fact unknown, we estimate it using maximum likelihood:

$$\hat{\Sigma}_F = \frac{1}{n} (Y - X \hat{\beta}_{q-1})' (Y - X \hat{\beta}_{q-1}), \tag{23}$$

where the subscript F stands for “full covariance,” and we use $\hat{\beta}_{q-1}$ instead of $\hat{\beta}_q$ to prevent overfitting, as in the single-response case of Section 1.2.1.

In practice, we find that estimating all h^2 entries of Σ leads to overfitting. Therefore, in our experiments, we estimate Σ using a diagonal matrix $\hat{\Sigma}_D$ that's the same as $\hat{\Sigma}_F$ except that the off-diagonal entries have been set to 0. In this case, (22) reduces to a sum of h terms like (14) for each response separately. We also experimented with shrunken estimates of the form $\hat{\Sigma}_\lambda = \lambda \hat{\Sigma}_D + (1 - \lambda) \hat{\Sigma}_F$ for $\lambda \in [0, 1]$ and observed informally that for $\lambda > 0.5$ or so, $\hat{\Sigma}_\lambda$ performed comparably with $\hat{\Sigma}_D$. However, there did not appear to be a performance advantage, so we continued to use $\hat{\Sigma}_D$, which is faster computationally.¹¹

2.2.3 Comparing the Coding Schemes

This section has discussed three information-theoretic approaches to multitask regression: RIC, Full MIC, and Partial MIC. In general, the negative log-likelihood portion of RIC may differ from that of the other two, because Full and Partial MIC may use a nondiagonal covariance estimate like $\hat{\Sigma}_F$, while RIC, only operating on one response at a time, implicitly uses $\hat{\Sigma}_D$. However, since we also use $\hat{\Sigma}_D$ for Full and Partial MIC, the real differences come from the coding penalties.

These are compared in Table 2 for various values of k , the number of responses for which we add the current feature under consideration. Of course, Full MIC is only allowed to take $k = 0$ or $k = h$, so it actually has h nonzero coefficients in all three rows of the table. However, if the extra $h - k$ coefficients correspond to non-predictive features, the extra reduction in residual-coding cost that Full MIC enjoys over the other methods is likely to be small.

The numbers beside the coding-cost formulas illustrate the case $m = 2,000$ features and $h = 20$ responses. As expected, each coding scheme is cheapest in the case for which it was designed; however, we note that the MIC methods are never excessively expensive, unlike RIC for $k = h$.

2.3 Implementation

As mentioned in Section 1.1, searching over all possible combinations of zero and nonzero coefficients for the model that minimizes description length is computationally intractable. We therefore implement MIC using a forward stepwise search procedure, detailed in Algorithm 2.1. Beginning with a null model containing only an intercept feature, we evaluate each feature for addition to this model and add the one that would most reduce description length, as computed using Algorithm

¹⁰We could also assume a uniform distribution on $\{1, 2, \dots, h\}$ and spend $\lg h$ bits to code k 's index. However, in practice we find that smaller values of k occur more often, so that the \lg^* -based code is generally cheaper.

¹¹We also experimented with two other regularized full covariance-matrix estimators, proposed in [LW04] and [SORS], respectively. While we never ran systematic comparisons, our informal assessment was that these methods generally did not improve performance relative to our current approach and sometimes reduced it.

Table 2: Costs in bits for each of the three schemes to code the appearance of a feature in $k = 1$, $k = \frac{h}{4}$, and $k = h$ response models. In general, we assume $m \gg h \gg 1$. Note that for $h \in \{5, \dots, 1000\}$, $c_h \approx 1$. Examples of these values for $m = 2,000$ and $h = 20$ appear in brackets; the smallest of the costs appears in bold.

k	Partial MIC	Full MIC	RIC
1	$\lg m + c_h + \lg h + 2$ [18.4]	$\lg m + 2h$ [51.0]	$\lg m + 2$ [13.0]
$\frac{h}{4}$	$\lg m + \lg^* \left(\frac{h}{4}\right) + c_h + \lg \binom{h}{h/4} + \frac{h}{2}$ [39.8]	$\lg m + 2h$ [51.0]	$\frac{h}{4} \lg m + \frac{h}{2}$ [64.8]
h	$\lg m + \lg^* h + c_h + 2h$ [59.7]	$\lg m + 2h$ [51.0]	$h \lg m + 2h$ [259.3]

2.2. Updating our model, we re-evaluate each remaining feature and add to the model the next-best feature. We continue until we find no more features that would reduce description length if added to the model.

```

Algorithm 2.1: STEPWISEREGRESSIONMIC( $X, Y$ , method)

// “method” is either “partially dependent MIC” or “fully dependent MIC.”
Make sure that the matrix  $X$  contains a column of 1’s for the intercept.
Initialize the model  $\hat{\beta}$  to have nonzero coefficients only for the intercept terms. That is, if  $\hat{\beta}$ 
is an  $m \times h$  matrix, exactly the first row has nonzero elements.
Compute  $\hat{\Sigma}$  based on the residuals  $Y - X\hat{\beta}$  (whether as  $\hat{\Sigma}_F$ ,  $\hat{\Sigma}_D$ , or something in between).
while true
{ Find the feature  $f$  that, if added, would most reduce description length.
  (If method is “fully dependent MIC,” then “adding a feature” means making
  nonzero the coefficients for all of the responses with that feature. If method is “partially
  dependent MIC,” then it means making nonzero the optimal subset of responses,
  where the “optimal subset” is approximated by a stepwise-style search
  through response subsets.)
  Let  $\hat{\beta}_f$  denote the model that would result if this new feature were added.
  if DESCRLENGTH( $X, Y, \hat{\beta}_f, \hat{\Sigma}$ , method) < DESCRLENGTH( $X, Y, \hat{\beta}, \hat{\Sigma}$ , method)
  then {  $\hat{\beta} \leftarrow \hat{\beta}_f$ 
        { Update  $\hat{\Sigma}$ .
  else break
return ( $\hat{\beta}$ )
  
```

```

Algorithm 2.2: DESCRLENGTH( $X, Y, \hat{\beta}, \hat{\Sigma}$ , method)

Let  $m$  be the number of features (columns of  $X$ ) and  $h$  the number of responses (columns of  $Y$ ).
 $d \leftarrow 0$  // Initialize  $d$ , the description length.
 $d \leftarrow d + \mathcal{D}(Y | \hat{\beta})$  // Add the likelihood cost using (22).
for  $j \leftarrow 1$  to  $m$  // Add the cost of feature  $j$ , if any.
{  $k \leftarrow$  number of nonzero responses for feature  $j$ 
  if  $k > 0$ 
  do {
    then {  $d \leftarrow d + \lg m$  // Cost to specify which feature.
          {  $d \leftarrow d + 2k$  // Cost to specify nonzero coefficients.
            if method == “partially dependent MIC”
            then  $d \leftarrow d + \lg^* k + c_h + \lg \binom{h}{k}$  // Specify which subset of responses.
          }
        }
  }
return ( $d$ )
  
```

The above procedure requires evaluating the quality of features $\mathcal{O}(mr)$ times, where r is the number of features eventually added to the model. For Partial MIC, each time we evaluate a feature, we have to determine the best subset of responses that might be associated with that feature. As Algorithm 2.1 notes, this is also done using a stepwise-style search: Start with

no responses, add the best response, then re-evaluate the remaining responses and add the next-best response, and so on.¹² Unlike an ordinary stepwise algorithm, however, we don't terminate the search if, at the current number of responses in the model, the description length fails to be the best seen yet. Because we're interested in borrowing strength across responses, we need to avoid overlooking cases where the correlation of a feature with any single response is insufficiently strong to warrant addition, yet the correlations with all of the responses are. Moreover, the $\lg \binom{h}{k}$ term in Partial MIC's coding cost does not increase monotonically with k , so even if adding the feature to an intermediate number of response models doesn't look promising, adding it to all of them might. Thus, we perform a full $\mathcal{O}(h^2)$ search in which we eventually add all the responses to the model. As a result, Partial MIC requires a total of $\mathcal{O}(mrh^2)$ evaluations of description length.¹³

However, a few optimizations can reduce the computational cost of Partial MIC in practice.

- We can quickly filter out most of the irrelevant features at each iteration by evaluating, for each feature, the decrease in negative log-likelihood that would result from simply adding it with all of its responses, without doing any subset search. Then we keep only the top t features according to this criterion, on which we proceed to do the full $\mathcal{O}(h^2)$ search over subsets. We use $t = 75$, but we find that as long as t is bigger than, say, 10 or 20, it makes essentially no impact to the quality of results. This reduces the number of model evaluations to $\mathcal{O}(mr + rth^2)$.
- We can often short-circuit the $\mathcal{O}(h^2)$ search over response subsets by noting that a model with more nonzero coefficients always has lower negative log-likelihood than one with fewer nonzero coefficients. This allows us to get a lower bound on the description length for the current feature for each number $k \in \{1, \dots, h\}$ of nonzero responses that we might choose as

$$\begin{aligned}
 & \text{(model cost for other features already in model)} \\
 & + \text{(negative log-likelihood of } Y \text{ if all } h \text{ responses for this feature were nonzero)} \\
 & + \text{(the increase in model cost if } k \text{ of the responses were nonzero)}.
 \end{aligned} \tag{24}$$

We then need only check those values of k for which (24) is smaller than the best description length for any candidate feature's best response subset seen so far. In practice, with $h = 20$, we find that evaluating k up to, say, 3 or 6 is usually enough; i.e., we typically only need to add 3 to 6 responses in a stepwise manner before stopping, with a cost of only $3h$ to $6h$.¹⁴

2.4 Experiments

We evaluate MIC on several synthetic and real data sets in which multiple responses are associated with the same set of features. We focus on the parameter regimes for which MIC was designed: Namely, $m \gg n$, but with only a relatively small number of features expected to be predictive. We describe the details of each data set in its own subsection below.

For comparing MIC against existing multitask methods, we used the Matlab "Transfer Learning Toolkit" [KR], which provides implementations of seven algorithms from the literature. Unfortunately, most of them did not apply to our data sets, since they often required *meta-features* (features describing other features), or expected the features to be frequency counts, or were unsupervised learners. The two methods that did apply were AndoZhang and BBLasso, both described in Section 2.1.

The AndoZhang implementation was written by the Transfer Learning Toolkit authors, Wei-Chun Kao and Alexander Rakhlin. It included several free parameters, including the regularization coefficients λ_k for each task (all set to 1, as was the default), the optimization method (also set to the default), and H . [AZ05, pp. 1838-39] tried values of H between 10 and 100, settling on 50 but finding that the exact value was generally not important. We performed informal cross-validation for values between 1 and 250 and found, perhaps surprisingly, that $H = 1$ consistently gave the best results. We used this value throughout the experiments below.

The BBLasso implementation was written by Guillaume Obozinski, based on a paper [OTJ06] published earlier than the [OTJ09] cited above; however, the algorithm is essentially the same. For each parameter, we used the default package setting.

AndoZhang and BBLasso are both classification methods, so in order to compare against them, we had to turn MIC into a classification method as well. One way would have been to update (22) to reflect a logistic model; however, the resulting computation of $\hat{\beta}$ would then have been nonlinear and much slower. Instead, we continue to apply (22) and simply regress on responses that have the values 0 or 1. Once we've chosen which features will enter which models, we do a final round of logistic regression for each response separately with only the chosen features.¹⁵

¹²A stepwise search that re-evaluates the quality of each response at each iteration is necessary because, if we take the covariance matrix $\hat{\Sigma}$ to be nondiagonal, the values of the residuals for one response may affect the likelihood of residuals for other responses. If we take $\hat{\Sigma}$ to be diagonal, as we end up doing in practice, then an $\mathcal{O}(h)$ search through the responses without re-evaluation would suffice.

¹³Full MIC, not requiring the search through subsets of responses, is only $\mathcal{O}(mr)$ in the number of evaluations of description length.

¹⁴If $\hat{\Sigma}$ is diagonal and we don't need to re-evaluate residual likelihoods at each iteration, the cost is only 3 to 6 evaluations of description length.

¹⁵This contrasts with BBLasso, which in the default Transfer Learning Toolkit implementation uses the original regression coefficients (mostly

2.4.1 Synthetic Data

We created synthetic data according to three separate scenarios, which we’ll call “Partial,” “Full,” and “Independent.” For each scenario, we generated a matrix of continuous responses as

$$Y_{\text{sim}} = X_{\text{sim}}\beta_{\text{sim}} + \epsilon_{\text{sim}},$$

with $m = 2,000$ features, $h = 20$ responses, and $n = 100$ observations. Then, to produce binary responses, we set to 1 those response values that were greater than or equal to the average value for their columns and set to 0 the rest, yielding a roughly 50-50 split between 1’s and 0’s because of the normality of the data. Each entry of X_{sim} was i.i.d. $\mathcal{N}(0, 1)$, each nonzero entry of β_{sim} was i.i.d. $\mathcal{N}(0, 1)$, and entry of ϵ_{sim} was i.i.d. $\mathcal{N}(0, 0.1)$, with no covariance among the ϵ_{sim} entries for different responses.¹⁶ Each response had 4 beneficial features, i.e., each column of β_{sim} had 4 nonzero entries.

The scenarios differed according to the distribution of the beneficial features in β_{sim} .

- In the Partial scenario, the first feature was shared across all 20 responses, the second was shared across the first 15 responses, the third across the first 10 responses, and the fourth across the first 5 responses. Because each response had four features, those responses (6 – 20) that didn’t have all of the first four features had other features randomly distributed among the remaining features (5, 6, . . . , 2000).
- In the Full scenario, each response shared exactly features 1 – 4, with none of features 5 – 2000 being part of the model.
- In the Independent scenario, each response had four random features among $\{1, \dots, 2000\}$.

Figure 1 illustrates these feature distributions, showing the first 40 rows of random β_{sim} matrices.

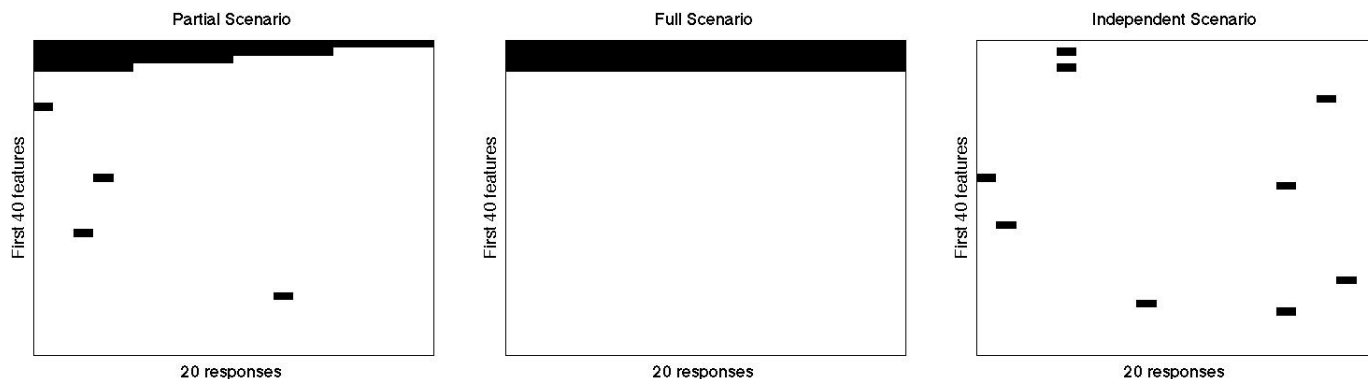


Figure 1: Examples of β_{sim} for the three scenarios, with the nonzero coefficients in black. The figures show all 20 columns of β_{sim} but only the first 40 rows. (For the Partial and Independent scenarios, the number of nonzero coefficients appearing in the first 40 rows is exaggerated above what would be expected by chance for illustration purposes.)

Table 3 shows the performance of each of the methods on five random instances of these data sets. Test-set errors for the “True Model” are those obtained with logistic regression for each response separately using a model containing exactly the features of β_{sim} that had nonzero coefficients. In addition to test-set error, we show precision and recall metrics. *Coefficient-level* precision and recall refer to individual coefficients: For those coefficients of the data-generating β_{sim} that were nonzero, did our final $\hat{\beta}$ show them as nonzero, and vice versa? *Feature-level* precision and recall look at the same question for entire features: If a row of β_{sim} had *any* nonzero coefficients, did our corresponding row of $\hat{\beta}$ have any nonzero coefficients (not necessarily the same ones), and vice versa?

The best test-set error values are bolded (though not necessarily statistically significant). As we would expect, each of Partial MIC, Full MIC, and RIC is the winner in the synthetic-data regime for which it was designed, although Partial MIC competes with Full MIC even in the Full regime. Though the results are not shown, we observed informally that AndoZhang and BBLasso tended to have larger differences between training and testing error than MIC, implying that MIC may be less

zeros). To enhance comparability, we should have adopted a consistent approach for evaluating test-set predictions: Either (1) use the original coefficients or (2) regress again on the selected features, for both MIC and BBLasso.

¹⁶We mentioned in Section 2.2.2 that MIC performed best when we used $\hat{\Sigma}_D$, the diagonal covariance-matrix estimate. One might wonder whether this was due only to the fact that we created our synthetic test data without covariance among the responses. However, this was not the case. When we generated synthetic noise using a correlation matrix in which each off-diagonal entry was a specific nonzero value (in particular, we tried 0.4 and 0.8), the same trend appeared: MIC with $\hat{\Sigma}_D$ had slightly lower test error.

Table 3: Test-set accuracy, precision, and recall of MIC and other methods on 5 instances of the synthetic data sets generated as described in Section 2.4.1. Because synthetic data is cheap and our algorithms, once trained, are fast to evaluate, we used 10,000 test data points for each data-set instance. Standard errors are reported over each task; that is, with 5 data sets and 20 tasks per data set, the standard errors represent the sample standard deviation of 100 values divided by $\sqrt{100}$ (except for feature-level results, which apply only to entire data sets and so are divided by $\sqrt{5}$). Baseline accuracy, corresponding to guessing the majority category, is roughly 0.5. AndoZhang’s NA values are due to the fact that it does not explicitly select features.

Method	True Model	Partial MIC	Full MIC	RIC	AndoZhang	BBLasso
Partial Synthetic Data Set						
Test error	0.07 ± 0.00	0.10 ± 0.00	0.17 ± 0.01	0.12 ± 0.01	0.50 ± 0.02	0.19 ± 0.01
Coeff. precision	1.00 ± 0.00	0.84 ± 0.02	0.26 ± 0.01	0.84 ± 0.02	NA	0.04 ± 0.00
Coeff. recall	1.00 ± 0.00	0.77 ± 0.02	0.71 ± 0.03	0.56 ± 0.02	NA	0.81 ± 0.02
Feature precision	1.00 ± 0.00	0.99 ± 0.01	0.97 ± 0.02	0.72 ± 0.05	NA	0.20 ± 0.03
Feature recall	1.00 ± 0.00	0.54 ± 0.05	0.32 ± 0.03	0.62 ± 0.04	NA	0.54 ± 0.01
Full Synthetic Data Set						
Test error	0.07 ± 0.00	0.08 ± 0.00	0.08 ± 0.00	0.11 ± 0.01	0.45 ± 0.02	0.09 ± 0.00
Coeff. precision	1.00 ± 0.00	0.98 ± 0.01	0.80 ± 0.00	0.86 ± 0.02	NA	0.33 ± 0.03
Coeff. recall	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.63 ± 0.02	NA	1.00 ± 0.00
Feature precision	1.00 ± 0.00	0.80 ± 0.00	0.80 ± 0.00	0.36 ± 0.06	NA	0.33 ± 0.17
Feature recall	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	NA	1.00 ± 0.00
Independent Synthetic Data Set						
Test error	0.07 ± 0.00	0.17 ± 0.01	0.36 ± 0.01	0.13 ± 0.01	0.49 ± 0.00	0.35 ± 0.01
Coeff. precision	1.00 ± 0.00	0.95 ± 0.01	0.06 ± 0.01	0.84 ± 0.02	NA	0.02 ± 0.00
Coeff. recall	1.00 ± 0.00	0.44 ± 0.02	0.15 ± 0.02	0.58 ± 0.02	NA	0.43 ± 0.02
Feature precision	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.83 ± 0.02	NA	0.30 ± 0.05
Feature recall	1.00 ± 0.00	0.44 ± 0.02	0.14 ± 0.02	0.58 ± 0.03	NA	0.42 ± 0.06

likely to overfit.¹⁷ Resistance to overfitting is a general property of MDL approaches, because the requirement of completely coding $\hat{\beta}$ constrains model complexity. This point can also be seen from the high precision of the information-theoretic methods, which, surprisingly, show comparable recall to BBLasso. Full MIC behaves most like BBLasso, because both algorithms select entire rows of coefficients for addition to $\hat{\beta}$.

2.4.2 Yeast Growth

Our first real data set comes from [LCCP09, pp. 5-6]. It consists of real-valued growth measurements of 104 strains of yeast ($n = 104$ observations) under 313 drug conditions. In order to make computations faster, we hierarchically clustered these 313 conditions into 20 groups using single-link clustering with correlation as the similarity measure. Taking the average of the values in each cluster produced $h = 20$ real-valued responses, which we then binarized into two categories: values at least as big as the average for that response (set to 1) and values below the average (set to 0).¹⁸ The features consisted of 526 markers (binary values indicating major or minor allele) and 6,189 transcript levels in rich media for a total of $m = 6,715$ features.

The “Yeast Growth” section of Table 4 shows test errors from 5-fold CV on this data set. Though the difference isn’t statistically significant, Partial MIC appears to outperform BBLasso on test error. In any event, Partial MIC produces a much sparser model, as can be seen from the numbers of nonzero coefficients and features: Partial MIC includes nonzero coefficients in an average of 4 rows of $\hat{\beta}$, in contrast to 63 for BBLasso. Partial MIC also appears to outperform Full MIC and RIC, presumably because the assumptions of complete sharing or no sharing of features across responses rarely hold for real-world data. Like Partial MIC, AndoZhang did well on this data set; however, the algorithm scales poorly to large numbers of responses and so took 39 days to run.¹⁹

¹⁷On the other hand, overfitting is a function of the regularization parameter chosen. We merely used the default parameter values in the Transfer Learning Toolkit, but we presumably could have controlled overfitting by using higher regularization penalties.

¹⁸The split was not exactly 50-50, as the data were sometimes skewed, with the mean falling above or below the median. Table 4 reflects this fact, showing that a classifier that simply guesses the majority label achieves an average error rate of 0.41.

¹⁹Much of the reason for this is that AndoZhang, as currently implemented, only produces predictions for one of the responses after it trains on all of them. Thus, in order to predict all 20 responses, we had to run it separately 20 times. It is probably possible to train once and then predict

Table 4: Accuracy and number of coefficients and features selected on five folds of CV for the Yeast Growth, Yeast Markers, and Breast Cancer data sets. Standard errors are over the five CV folds; i.e., they represent (sample standard deviation) / $\sqrt{5}$. The “Majority Label” column represents a classifier which guesses the more common of the 0 / 1 labels seen in the training set. AndoZhang’s NA values are due to the fact that it does not explicitly select features.

Method	Partial MIC	Full MIC	RIC	AndoZhang	BBLasso	Majority Label
Yeast Growth						
Test error	0.38 ± 0.04	0.39 ± 0.04	0.41 ± 0.05	0.39 ± 0.03	0.43 ± 0.03	0.41 ± 0.04
Num. coeff. sel.	22 ± 4	64 ± 4	9 ± 1	NA	1268 ± 279	NA
Num. feat. sel.	4 ± 0	3 ± 0	9 ± 1	NA	63 ± 14	NA
Yeast Markers						
Test Error	0.34 ± 0.07	0.44 ± 0.04	0.27 ± 0.06	-	0.45 ± 0.05	0.53 ± 0.03
Num. coeff. sel.	20 ± 1	68 ± 26	37 ± 2	NA	1044 ± 355	NA
Num. feat. sel.	16 ± 1	3 ± 1	37 ± 2	NA	52 ± 18	NA
Breast Cancer						
Test error	0.33 ± 0.08	0.37 ± 0.08	0.36 ± 0.08	0.44 ± 0.03	0.33 ± 0.08	0.39 ± 0.04
Num. coeff. sel.	3 ± 0	11 ± 1	2 ± 0	NA	61 ± 19	NA
Num. feat. sel.	2 ± 0	2 ± 0	2 ± 0	NA	12 ± 4	NA

2.4.3 Yeast Markers

The Yeast Growth data set described above includes as features both markers and transcripts for 104 yeast strains. We can consider these as variables for prediction in their own right, without reference to the growth responses at all. In fact, regression of transcripts against markers is commonly done; it’s known as “expression quantitative trait loci” (eQTL) mapping [KW06]. Since the marker variables are already binary (major or minor allele at the given location), we decided to flip the problem around from the usual eQTL setup: Using the $m = 6,189$ transcripts as features, we predicted a subset of 20 of the 526 markers (numbered 25, 50, 75, . . . , 500 for a good distribution along the genome).

The results are shown in the “Yeast Markers” section of Table 4. Unlike the case of the Yeast Growth data, there is apparently less sharing of feature information across markers, as RIC appears to outperform Partial MIC on test error. We did not run AndoZhang on this data set.

2.4.4 Breast Cancer

Our Breast Cancer data set represents a combination of five of the seven data sets used in [vtVDvdV+02]. It contains 1,171 observations for 22,268 RMA-normalized gene-expression values. We considered five associated responses; two were binary—prognosis (“good” or “poor”) and ER status (“positive” or “negative”)—and three were not—age (in years), tumor size (in mm), and grade (1, 2, or 3). We binarized the three non-binary responses into two categories: Response values at least as high as the average, and values below the average. Some of the responses were unavailable for some observations, so we eliminated those observations, leaving 882. Of those, we kept only the first $n = 100$, both to save computational resources and to make the problem “harder.” To reduce the features to a manageable number, we took the $m = 5,000$ that had highest variance.

Table 4 shows the results. In this case, BBLasso did as well as Partial MIC on test error; however, as usual, Partial MIC produced a much sparser model. Sparsity is important for biologists who want to interpret the selected features.

3 Hypothesis Testing with a Single Response

As the experimental results in Section 2.4 demonstrate, regression can be an effective tool for predicting one or more responses from features, and in cases where the number of features is large, selection can improve generalization performance. However, as we noted with the Yeast and Breast Cancer data sets, accuracy is not always the only goal; sometimes model interpretability is important. Indeed, in science, the entire goal of regression is often not prediction but rather discovering a “true model,” by testing hypotheses about whether pairs of variables really have a linear relationship. For instance, an econometrician regressing housing prices on air-pollution levels and resident income [HR78] is not trying to predict price levels but to study consumer demand behavior.

all 20 responses simultaneously, but we wanted to avoid introducing errors by changing the code.

Hypothesis testing with a small number of features is straightforward: Given a regression coefficient, we can compute a t statistic, and if it exceeds some threshold, we reject the null hypothesis that the coefficient is 0. The situation becomes slightly more complicated when we have many features whose coefficients we want to examine. Section 3.1 reviews some standard approaches to this multiple-testing problem, and Section 3.2 recasts them in the light of MDL. This sets the stage for an MIC approach to hypothesis testing in Section 4.

3.1 Multiple-Testing Procedures

Consider the case of a single response variable y (say, a gene-transcript level), and suppose we're trying to determine which of m features (say, genetic markers) are linearly correlated with the response. If we test each feature at a significance level α , the overall probability that we'll falsely reject a true null hypothesis will be much greater than α —this is the problem of *multiple hypothesis testing*.

3.1.1 Bonferroni Correction

A standard way to control against so-called *alpha inflation* is called the *Bonferroni correction*: Letting H_1, \dots, H_m denote the null hypotheses and p_1, \dots, p_m the associated p-values, we reject H_j when $p_j \leq \frac{\alpha}{m}$. By Boole's inequality, this controls what is known as the *family-wise error rate* (FWER), the probability of making any false rejection, at level α under the *complete null hypothesis* that all of H_1, \dots, H_m are true.²⁰ In fact, Bonferroni controls FWER in a *strong sense* as well: For any subset of null hypotheses, the probability of falsely rejecting any member of that subset when all members are true is also bounded by α [Hoc88, p. 800].

The Bonferroni correction is a *single-stage testing procedure* in which, unfortunately, testing a larger number of hypotheses reduces the power for rejecting any one of them [Sha95, p. 569]. This is especially problematic when the number of hypotheses tested is in the thousands or tens of thousands, as is common with, e.g., microarray or fMRI data. However, statisticians have developed several *multistage testing procedures* that help to overcome this limitation by conditioning rejection thresholds for some test statistics on the results of rejections by others.

3.1.2 Improvements on Bonferroni

One of the first of these was the *Holm step-down procedure* [Hol79]. Here, we let $p_{(1)}, \dots, p_{(m)}$ denote the p-values sorted in increasing order and $H_{(1)}, \dots, H_{(m)}$ the corresponding null hypotheses. We begin by looking at $p_{(1)}$: If it fails to be $\leq \frac{\alpha}{m}$, we stop without rejecting anything. (This is what Bonferroni would do as well, since no p-value is $\leq \frac{\alpha}{m}$.) However, if we do reject $H_{(1)}$, we move on to $H_{(2)}$ and reject if and only if $p_{(2)} \leq \frac{\alpha}{m-1}$. We continue in this manner, rejecting $H_{(j)}$ when $p_{(j)} \leq \frac{\alpha}{m-j+1}$, until we fail to reject a hypothesis. Like Bonferroni, the Holm method controls FWER in the strong sense for independent or dependent test statistics [Hoc88, p. 800].

Simes [Sim86, p. 751] proposed a more lenient approach: Reject the complete null hypothesis if *any* of the following inequalities holds, $j = 1, 2, \dots, m$:

$$p_{(j)} < \frac{j\alpha}{m}.$$

This controls FWER for independent test statistics at level α . Moreover, simulation studies suggested that it remained true for various multivariate-normal and multivariate-gamma test statistics [Sim86, p. 752]. Unfortunately, unlike the Bonferroni and Holm procedures, the Simes approach says nothing about rejecting individual hypotheses once the complete null is rejected [Sim86, p. 754], although [Hoc88] and [Hom88] subsequently proposed limited procedures for doing so.

3.1.3 FDR and the BH Procedure

In his closing discussion, Simes [Sim86, p. 754] proposed that if we wanted to reject individual null hypotheses, we might reject the ordered hypotheses $H_{(1)}, \dots, H_{(q)}$ such that

$$q = \max \left\{ j : p_{(j)} \leq \frac{j\alpha}{m} \right\}. \tag{25}$$

However, he cautioned that there was “no formal basis” for this. Indeed, [Hom88, p. 384] showed that in certain cases, the FWER of this procedure could be made arbitrarily close to 1 for large m .

²⁰When the test statistics are independent or *positive orthant dependent*, one can replace $\frac{\alpha}{m}$ by $1 - (1 - \alpha)^{\frac{1}{m}}$, but the resulting improvement in power is small for small α [Sha95, p. 570].

However, Benjamini and Hochberg [BH95] pointed out that the Simes procedure could be said to control a different measure, which they called the *false-discovery rate* (FDR). Letting V denote the (unobservable) random variable for the number of true null hypotheses rejected and S the (unobservable) random variable for the number of correctly rejected null hypotheses,

$$\text{FDR} := P(V + S > 0) E \left[\frac{V}{V + S} \mid V + S > 0 \right]$$

(where $V + S$, the number of null hypotheses rejected in total, is observable). Thus, FDR is the expected proportion of falsely rejected null hypotheses. This statistic is more flexible than FWER because it accounts for the total number of hypotheses considered; if, for instance, we had $m = 1,000,000$ hypotheses, the FWER would be very high, but the *proportion* of false rejections could still be low. Because of this flexibility for large m , FDR has become standard in bioinformatics, neuroscience, and many other fields.

[BH95, p. 293] showed that the *Benjamini-Hochberg (BH) step-up procedure* (25) controls FDR at α for independent test statistics and any configuration of false null hypotheses. [BY01] extended this result to show that the procedure also controlled FDR for certain types of positive dependency. This included positively correlated and normally distributed one-sided test statistics, which occur often in, for instance, gene-expression measurements [RYB03, p. 370].

Various extensions to the BH procedure have been proposed. For instance, [BL99] suggested a step-down approach for independent test statistics that, while not dominating the BH step-up procedure, tended experimentally to yield higher power. [BY01, p. 1169] showed that replacing α by $\alpha / \sum_{j=1}^m \frac{1}{j}$ would control FDR at α for any test-statistic correlation structure, though this is often more conservative than necessary (p. 1183). [YB99] proposed a resampling approach for estimating FDR distributions based the data, in order to gain increased power when the test statistics were highly correlated. Many further modifications to the BH procedure have been put forward since. However, in this thesis, we stick with the original version given by (25).

3.2 Hypothesis Testing by MDL

MDL compares log-likelihoods of data under various models, choosing a more complicated model if and only if it has sufficiently lower negative log-likelihood. Appendix A shows that, in fact, this process is equivalent to a standard statistical procedure known as a generalized likelihood-ratio test, at some implied significance level α .

Hypothesis testing differs slightly from regression, however. The goal with regression is to minimize test-set error by choosing a few highly informative features. Because MIC penalizes for each feature, if we find several, very similar features that are all correlated with the response, stepwise regression by MIC will likely include only one of them. For instance, suppose the true model is $y = f_1 + 2f_2 + \epsilon$, where f_1 and f_2 are nearly identical features. Not wanting to waste model-coding bits, regression MIC would probably give the model $y = 3f_1 + \epsilon$ or $y = 3f_2 + \epsilon$.

With hypothesis testing, we aim to find all of the features that are significantly correlated with the response. In the above example, we would hope to include both f_1 and f_2 in our set of relevant features. To do this, we regress Y on each feature in isolation; we keep those features that, according to MDL, deserve nonzero regression coefficients. Conceptually, we can think of this as a for-loop over features, in a given iteration of which we call an MIC-stepwise-regression function using the regular Y matrix but with an X matrix that contains only the current feature (and an intercept). There's one catch, though: Since our for-loop searches through m potential features to add, we're effectively doing m hypothesis tests, and we need to incorporate some sort of multiple-testing penalty. Below we describe a way to motivate such a penalty from an MDL perspective.

Section 1.2 described a scenario to motivate ordinary MDL regression: Sender wanted to transmit data Y to Receiver when both Sender and Receiver knew the value of an associated matrix X of features. Now suppose instead that there are m different Receivers, where Receiver i knows the values only of the i^{th} feature (i.e., the i^{th} column of X). All the Receivers want to reconstruct Y , so Sender has to transmit messages to each of them telling them how to do it, possibly using the feature that they individually know. The messenger Hermes visits Sender and tells her some ground rules: Sender must transmit directly to each Receiver only information about how to reconstruct Y given a model. All model information itself (i.e., all the $\hat{\beta}$ coefficients) Sender has to transmit to Hermes, who will then visit each Receiver and tell him his appropriate coefficient (possibly 0 if Sender hasn't bothered to encode a model for that feature). As a bonus, Sender is allowed to include in each model the intercept coefficient of Y for free, meaning that if Sender doesn't specify a feature regression coefficient for Receiver i , that Receiver at least gets the average \bar{Y} to use in reconstructing Y . Letting $\hat{\beta}_q$ denote the model Sender tells Hermes, Sender's total description length is

$$\mathcal{D}(\hat{\beta}_q) + \mathcal{D}^m(Y | \hat{\beta}_q),$$

where

$$\mathcal{D}^m(Y | \hat{\beta}_q) := \sum_{i=1}^m \mathcal{D}(Y | \mathcal{M}_i). \quad (26)$$

Here, we define \mathcal{M}_i to be a null model, containing only the free intercept term, if Sender doesn't specify a nonzero regression coefficient for Receiver i in $\hat{\beta}_q$. Otherwise, \mathcal{M}_i contains both a costless intercept term and the regression coefficient for feature i .

The result should be something like a series of hypothesis tests, in which the features in $\hat{\beta}_q$ correspond to the rejected null hypotheses. The idea is that, because Sender has to tell Hermes which coefficients go to which Receiver, the coding of $\hat{\beta}_q$ will involve a penalty that grows with the number of possible features, which is what we need to protect against alpha inflation from multiple tests.

3.3 MDL Multiple-Testing Corrections

We consider two coding schemes that Sender might use to tell Hermes which features are contained in her model $\hat{\beta}_q$.

3.3.1 Bonferroni-Style Penalty

Suppose $m = 5,000$ and Sender wants to use nonzero coefficients for features 226, 1,117 and 3,486. One way to tell Hermes this is to transmit the binary representation of each of these numbers, using an idealized $\lg m$ bits for each one. Then, as in Section 1.2.2, Sender spends 2 bits to specify each coefficient. The resulting message to Hermes costs $3 \lg m + 3 \cdot 2$ bits. In general, Sender's description length will be

$$q \lg m + 2q + \mathcal{D}^m(Y | \hat{\beta}_q). \quad (27)$$

For each feature j , let $-\lg \Lambda_j$ denote the decrease in residual coding cost that would result from using a nonzero coefficient for feature j . (See the Appendix for the motivation behind this notation.) Then minimizing (27) is equivalent to the following decision rule: Looking at the features in decreasing order of their $-\lg \Lambda_j$ values, add feature j to the model as long as

$$-\lg \Lambda_j \geq \lg m + 2. \quad (28)$$

Interestingly, (39) in Appendix A.4 shows that this is equivalent to a Bonferroni correction at some implied α . In fact, according to Appendix A.5, taking a cost per coefficient around 2,²¹ as we did for stepwise regression, corresponds to an α around 0.05.

3.3.2 BH-Style Penalty

If Sender expects to code more than one coefficient, rather than coding an entire index to specify each feature, she may find it advantageous to use a scheme similar to (21). There, the scheme was used to convey a subset of nonzero response coefficients in Partial MIC; here it would describe a subset of features from $\{1, \dots, m\}$. Using this code, Sender's description length using $\hat{\beta}_q$ would be

$$\lg^* q + c_m + \lg \binom{m}{q} + 2q + \mathcal{D}^m(Y | \hat{\beta}_q). \quad (29)$$

Apart from the two (small) terms at the beginning, this can be made to correspond at some implied α ²² with (44) and thus approximately with the ordinary BH procedure.

4 MIC for Hypothesis Testing

As was the case with regression, real-world hypothesis-testing problems often involve multiple responses. [TVW05, sec. 1] discuss the task of identifying a subset of 770 wavelength features that best predict 14 chemical quantities (e.g., pH, organic carbon, and total cations). *Expression quantitative trait loci* (eQTL) mapping is the process of looking for correlations between potentially thousands of gene-expression transcripts and genetic markers (e.g., [SMD+03]). Section 4.1 reviews some existing ways to deal with this problem, and then Section 4.2 describes the MIC approach. Experimental comparisons of the various methods are described in Section 4.3.

²¹2.77 to be exact, for $m = 1$ and one degree of freedom, though this changes somewhat with m and the number of degrees of freedom.

²²In fact, the same α as for the Bonferroni case above.

4.1 Existing Approaches

We describe a few examples of ways to perform hypothesis testing with multiple responses.

4.1.1 Bonferroni and BH

Probably the most straightforward approach is to treat each response separately and apply, say, the Bonferroni or BH procedures of Section 3.1 one response at a time. We’ll refer to these approaches as simply “Bonferroni” and “BH,” respectively.

According to [KCY+06, p. 20], a number of early eQTL studies took this approach, applying single-transcript mapping methods to each transcript separately. However, because such methods are designed to control false positives when analyzing a single transcript, multiple tests across transcripts may result in inflated FDR [KCY+06, p. 23].

4.1.2 BonferroniMatrix and BHMatrix

To account for the greater number of hypothesis tests being performed with $h > 1$ responses, we might penalize not just by $\frac{\alpha}{m}$ but $\frac{\alpha}{mh}$. The procedure that consists in applying the latter Bonferroni threshold to each p-value in our $m \times h$ matrix of p-values is what we’ll call “BonferroniMatrix.” Effectively, we’re imagining our matrix of p-values as one long vector and applying the standard Bonferroni correction to that.

Similarly, we can imagine turning our matrix of p-values into a single vector and applying the BH method to it: We’ll call this approach “BHMatrix.” In contrast to BH, it has a harsher starting penalty ($\frac{\alpha}{mh}$ rather than $\frac{\alpha}{m}$), but it also allows us to pick out the best p-values from any location in the entire matrix, not just p-values in the particular column (response) that we’re currently examining. [KCY+06, p. 21] describes a conceptually similar approach, called “Q-ALL,” that used so-called q-values [Sto03] to identify significant marker-transcript correlations.

The result of BonferroniMatrix at level α is the same as that of regular Bonferroni at level $\frac{\alpha}{h}$, so if we’re looking at a range of significance levels, it suffices to examine just Bonferroni or BonferroniMatrix; we’ll show results for Bonferroni only in what follows. The same is not true for BH vs. BHMatrix, so we report on both methods.

4.1.3 Empirical Bayes

The BHMatrix approach does not treat each response separately, since, for example, if one response has lots of low p-values that pass the harshest thresholds of the step-up procedure, it can leave easier thresholds for the other responses. But this doesn’t really “share information” across responses in a meaningful way.

One approach that does is called *EBarrays* [KNLG03], which uses an empirical Bayes hierarchical model to estimate a prior for the underlying means of each transcript (response), allowing for stable inference across transcripts despite the small sample size of each individually. The goal is to determine, at each marker (feature), which transcripts show significant association. Originally designed for sharing transcript information one marker at a time, the method has been extended to a *mixture over markers* (MOM) model that uses the EM algorithm to assign probabilities that a transcript correlates to each marker [KCY+06]. As the names suggest, these methods apply specifically to eQTL problems, though of course, the empirical-Bayes shrinkage framework applies more generally. We give this as just an example of previous approaches to sharing strength across responses that have been developed for hypothesis testing.

4.2 The MIC Approach

While the EBarrays approach shares information across responses, it fails to take advantage of one potentially important source of shared strength: Namely, that features strongly associated with one response are perhaps more likely to be associated with other responses. In the case of eQTL, for example, we might suspect this type of sharing across responses by marker features corresponding to *trans*-regulatory elements, which affect expression of many different genes.

On the other hand, the MIC approach of Section 2 does pick up on sharing of features across responses, so we propose a method for hypothesis testing by MIC. Just as the single-response version of hypothesis testing by MDL in Section 3.2 applied single-response regression to each feature individually, so we can also imagine applying the multiple-response MIC regression approach of Section 2.2 to each feature individually in order to do multiple-response hypothesis testing.

This amounts basically to a for-loop of the MIC Stepwise Regression Algorithm 2.1 over each feature, except that the costs of coding the features are different from usual because of the setup that Hermes imposed (see Section 3.2). Just as in that section, we can consider a Bonferroni-style (“Bonferroni-MIC”) and a BH-style (“BH-MIC”) coding scheme, using feature penalties of $q \lg m$ and $\lg^* q + c_m + \lg \binom{m}{q}$, respectively. Below we make each of these approaches more explicit.

4.2.1 Implementation

With Bonferroni-MIC, we can evaluate each feature one at a time for inclusion based on whether `STEPWISEREGRESSIONMIC` (Algorithm 2.1) would have included nonzero coefficients for that feature, with the following exception. That algorithm includes a feature penalty of \lg of the number of features (columns) in the given X matrix. Here, when we give the algorithm only one column of X at a time, this term would normally be just $\lg 1$. However, the coding scheme that Hermes imposed requires us instead to use $\lg m$ bits to specify a feature, where m is the total number of features in X . Pseudocode appears in Algorithm 4.1.

Algorithm 4.1: `BONFERRONI-MIC(X, Y, method)`

```

 $\hat{\beta} \leftarrow 0$ 
for  $i = 1$  to  $m$  // Loop over the features.
  { Essentially, we call
     $\hat{\beta}_i \leftarrow \text{STEPWISEREGRESSIONMIC}(X_i, Y, \text{method})$ ,
    except that the usual penalty term corresponding to  $\lg$  of the number of columns of the
    input  $X_i$  matrix (just 1 here) is replaced by  $\lg m$ , where  $m$  is the number of columns of  $X$ .
  }
return  $(\hat{\beta})$ 

```

BH-MIC is slightly trickier, because the cost of adding a new feature depends on how many features are already in the model. Our approach, outlined in Algorithm 4.2, is to loop through features and evaluate whether they would include nonzero coefficients when no feature-coding cost is imposed (i.e., no $(\lg m)$ -like term is charged). This fills $\hat{\beta}$ with a possibly inflated number of nonzero coefficients that we subsequently trim down if necessary. To do this, we evaluate the cost of keeping the best q of the currently nonzero features for each q from 0 to the current number of nonzero features q_{orig} , choose the optimal value q^* , and zero out the rest. Note that zeroing out those rows of $\hat{\beta}$ is sufficient; we have no need to go back and recompute the optimal subset of responses for the remaining features, because the cost to specify the included features is the same regardless of the response subset.

Algorithm 4.2: `BH-MIC(X, Y, method)`

```

 $\hat{\beta} \leftarrow 0$ 
for  $i = 1$  to  $m$  // Loop over the features.
  { Essentially, we call
     $\hat{\beta}_i \leftarrow \text{STEPWISEREGRESSIONMIC}(X_i, Y, \text{method})$ ,
    but without any cost to code features (the usual  $\lg m$  term). We also record  $s_i$ ,
    the (positive) number of bits saved by using the current subset of nonzero coefficients  $\hat{\beta}_i$ 
    instead of having all the coefficients zero. If  $\hat{\beta}_i == 0_{1 \times h}$ , i.e., all the coefficients
    are already all zeros,  $s_i \leftarrow 0$ .
  }
  Sort the  $s_i$  in decreasing order and call them  $s_{(i)}$ .  $s_{(1)}$  corresponds to saving the most bits, etc.
   $q_{\text{orig}} \leftarrow$  number of nonzero features currently in  $\hat{\beta}$  (possibly too many)
   $q^* \leftarrow \max_{q \in \{1, \dots, q_{\text{orig}}\}} \left\{ \sum_{i=1}^q s_{(i)} - \left( \lg^* q + c_m + \lg \binom{m}{q} \right) \right\}$  (or  $q^* \leftarrow 0$  if none of these is positive)
  Zero out the rows of  $\hat{\beta}$  corresponding to features whose inclusion saves fewer than  $s_{(q^*)}$  bits.
return  $(\hat{\beta})$ 

```

In both Bonferroni-MIC and BH-MIC, the algorithmic complexity is dominated by the for-loop. In Section 2.3, we saw that a call to `STEPWISEREGRESSIONMIC` required $\mathcal{O}(m_i r h^2)$ evaluations of description length, where we've used m_i to denote the number of columns of X_i . In particular, $m_i = 1$ for each i and r is at most 1, so the complexity within a given for-loop iteration is only $\mathcal{O}(h^2)$. Overall, then, these algorithms evaluate description length $\mathcal{O}(m h^2)$ times.

4.2.2 Hypothesis-Testing Interpretation

Consider a single feature. In Section 3.2, we noted the correspondence between MDL regression on that feature and a statistical likelihood-ratio test of whether its regression coefficient was nonzero. With $h > 1$ responses, MIC hypothesis

testing searches over subsets of responses to have nonzero coefficients with the feature. Each evaluation of description length during this process can be interpreted as a hypothesis test, in the way explained in the Appendix. In particular, when we evaluate the description length for some subset $S \subset \{1, 2, \dots, h\}$ of the responses being nonzero, we’re implicitly doing a likelihood-ratio test with these null and alternative hypotheses:

$$\begin{aligned} H_0 : \beta &= 0_{1 \times h} \quad \text{vs.} \\ H_1 : \beta_i &\neq 0 \text{ for } i \in S, \end{aligned} \tag{30}$$

where β is the $1 \times h$ matrix of true regression coefficients of the feature with the h responses. In theory, MDL performs exponentially many such tests, one for each subset S ,²³ though they are obviously highly correlated. We can interpret the $\lg^* k + c_h + \lg \binom{h}{k}$ penalty to specify the subset of k of the responses (as described in Section 2.2.1) as something of a multiple-testing correction for doing all of these implicit hypothesis tests.

Section 3.3 and the corresponding portions of the Appendix pointed out the approximate correspondence between particular penalties and particular corrections: For instance, that $\lg m$ in log-likelihood space is essentially a Bonferroni correction in p-value space, or that a cost per coefficient of 2.77 approximately corresponds to $\alpha = 0.05$. With $h > 1$ responses, these approximations are somewhat rougher. The reason is that in (30), $k = |S|$, the difference in dimension of the parameter spaces between H_0 and H_1 , is often greater than 1. In this case, $-2 \ln \Lambda \sim \chi_{(k)}^2$, and for $k > 1$, the standard-normal approximation using Φ in Appendix A.3 no longer goes through. Nevertheless, we can expect the Bonferroni- and BH-style penalties from Section 3.3 to do roughly the right things for multiple responses; we needn’t make them match the Bonferroni and BH procedures exactly. (If we did, what would be the point of using information theory?)

4.3 Experiments

When we measure the performance of our algorithms in terms of test-set error, we can use cross-validation to run experiments on real data sets, as we did in Section 2.4. When, instead, we need to judge how well our algorithm identifies truly nonzero coefficients, we have to fall back to synthetic data. That’s because the real world doesn’t tell us the true β matrix²⁴—we only know it if we make it ourselves. Thus, in this section, we rely entirely on synthetic data. However, we can sometimes base our synthetic data on a real-world data set, and we do this with the Yeast Growth data in Section 4.3.2.

For each of the scenarios below, we generated 25 random instances of the data sets, corresponding to 25 random β_{sim} matrices, taking $n = 100$ training data points and $h = 20$ responses. Because synthetic data is cheap, we evaluate test-set error on 10,000 test data points. We calculate error for each response as SSE/\sqrt{n} , where SSE is the sum of squares error from the ordinary-least-squares regression of the given response on exactly the features selected by the algorithm (plus an intercept). We report precision and recall at the coefficient level (whether each particular nonzero coefficient was selected). We calculate these results separately for each response in each data set, so that standard errors represent standard deviations divided by $\sqrt{20 \cdot 25}$.

We compare the following feature-selection approaches:

- “Truth”: Use an oracle to select exactly the true features.
- “Bonf, $\alpha=\alpha_0$ ”: For each response separately, select those features whose p-values with the response are $\leq \alpha_0/m$. We calculate the p-value for a given feature and response by regressing the response against only that feature and an intercept, and evaluating the p-value of the slope regression coefficient.
- “Indep,cpc= c_0 ”: Apply the RIC algorithm described in Section 2.2 to each response separately, using c_0 as the cost to code a coefficient value. (c_0 does not include the cost to specify which features enter the model, which is always $\lg m$. It shouldn’t fall below 0, and the lowest value we tried was 0.1.)
- “Bonf-MIC”: Bonferroni-style MIC, as described in Algorithm 4.1. The cost to code a coefficient is 2 bits.
- “BH, $\alpha=\alpha_0$ ”: For each response separately, calculate p-values the same way as with “Bonf, $\alpha=\alpha_0$,” but apply the BH procedure (25) at level α_0 instead of the Bonferroni correction.
- “BHMat, $\alpha=\alpha_0$ ”: Apply the BH procedure at level α_0 to the entire $m \times h$ matrix of p-values all at once.
- “BH-MIC”: BH-style MIC, as described in Algorithm 4.2. The cost to code a coefficient is 2 bits.

²³Of course, in practice we do only $\mathcal{O}(h^2)$ of them during our stepwise-style search through responses to include.

²⁴Indeed, there may not be any true β matrix, because the data are not *really* generated according to our model (18), though the approximation works well enough.

The MIC methods have no free parameters (if we fix the cost per coefficient at 2 bits), so they give point values for precision and recall. To allow for comparison, then, we ran the other methods at a variety of α_0 or c_0 levels, presenting results for α_0 or c_0 levels at which precision approximately matched that of MIC; relative performance can then be assessed based on recall. Because we tried only a discrete grid of α_0 or c_0 values, precisions did not always match exactly, so we took the highest precision not exceeding the MIC value. This puts MIC at a slight disadvantage in comparing recall, because its precision is often slightly higher than that of the other methods.

In our tables below, we make bold those results that represent the best performance among the three Bonferroni-style methods and also among the three BH-style methods. Here, “best performance” is based on the observed average value, but the differences are often not statistically significant.

4.3.1 Simulated Synthetic Data

We created three data sets in the same manner as described in Section 2.4.1, the only difference being that we took $m = 1,000$ features instead of 2,000 (for no particular reason). Table 5 shows the results for the Partial, Full, and Independent scenarios.

MIC appears to have had worse test-error than the other methods except in the Full scenario, where it performed close to the Truth. As the especially poor Independent performance suggests, this is probably due to the fact that MIC has a harder time picking up on single nonzero coefficients in a given row of β_{sim} , due to its “higher overhead” cost to put them in (see Table 2 for $k = 1$).

Of course, for hypothesis-testing algorithms, test error is not the most important metric. On recall, MIC does outperform its competitors in the Partial and especially Full scenarios. And as we would expect, it performs worse in the Independent case, again due to its high overhead for individual coefficients on a row of $\hat{\beta}$.

Partial Scenario							
Method	Truth	Bonf, $\alpha=1$	Indep,cpc=0.1	Bonf-MIC	BH, $\alpha=0.3$	BHMat, $\alpha=0.3$	BH-MIC
Train Err	0.31 \pm 0.00	0.52 \pm 0.01	0.63 \pm 0.01	0.56 \pm 0.01	0.51 \pm 0.01	0.52 \pm 0.01	0.53 \pm 0.01
Test Err	0.32 \pm 0.00	0.57 \pm 0.01	0.66 \pm 0.01	0.59 \pm 0.02	0.56 \pm 0.01	0.57 \pm 0.01	0.57 \pm 0.01
Coeff Prec	1.00 \pm 0.00	0.74 \pm 0.01	0.96 \pm 0.00	0.76 \pm 0.01	0.70 \pm 0.01	0.73 \pm 0.01	0.74 \pm 0.01
Coeff Rec	1.00 \pm 0.00	0.60 \pm 0.01	0.53 \pm 0.01	0.71 \pm 0.01	0.61 \pm 0.01	0.60 \pm 0.01	0.73 \pm 0.01
Full Scenario							
Method	Truth	Bonf, $\alpha=2$	Indep,cpc=0.1	Bonf-MIC	BH, $\alpha=0.45$	BHMat, $\alpha=0.5$	BH-MIC
Train Err	0.31 \pm 0.00	0.49 \pm 0.01	0.61 \pm 0.01	0.30 \pm 0.00	0.49 \pm 0.01	0.48 \pm 0.01	0.30 \pm 0.00
Test Err	0.32 \pm 0.00	0.54 \pm 0.01	0.63 \pm 0.01	0.33 \pm 0.00	0.54 \pm 0.01	0.53 \pm 0.01	0.33 \pm 0.00
Coeff Prec	1.00 \pm 0.00	0.62 \pm 0.01	0.97 \pm 0.00	0.62 \pm 0.01	0.59 \pm 0.01	0.57 \pm 0.01	0.61 \pm 0.01
Coeff Rec	1.00 \pm 0.00	0.61 \pm 0.01	0.53 \pm 0.01	0.99 \pm 0.00	0.61 \pm 0.01	0.61 \pm 0.01	0.99 \pm 0.00
Independent Scenario							
Method	Truth	Bonf, $\alpha=0.1$	Indep,cpc=0.1	Bonf-MIC	BH, $\alpha=0.04$	BHMat, $\alpha=0.05$	BH-MIC
Train Err	0.31 \pm 0.00	0.59 \pm 0.01	0.59 \pm 0.01	0.82 \pm 0.02	0.59 \pm 0.01	0.59 \pm 0.01	0.70 \pm 0.01
Test Err	0.32 \pm 0.00	0.62 \pm 0.01	0.62 \pm 0.01	0.86 \pm 0.02	0.62 \pm 0.01	0.62 \pm 0.01	0.73 \pm 0.02
Coeff Prec	1.00 \pm 0.00	0.96 \pm 0.01	0.96 \pm 0.01	0.96 \pm 0.01	0.96 \pm 0.01	0.96 \pm 0.01	0.96 \pm 0.01
Coeff Rec	1.00 \pm 0.00	0.53 \pm 0.01	0.54 \pm 0.01	0.40 \pm 0.01	0.54 \pm 0.01	0.54 \pm 0.01	0.47 \pm 0.01

Table 5: Test-set accuracy, precision, and recall of MIC and other methods on 25 instances of the synthetic data sets generated as described in Section 2.4.1, except with $m = 1,000$.

4.3.2 Simulated Yeast Data

We created a synthetic Yeast Growth data set based on the one described in Section 2.4.2. As will be explained below, talking about the “true coefficients” of a synthetic model requires that the features be relatively uncorrelated. However, the transcript features of the original data set were highly correlated, so we included only the 526 marker features in the X matrix. As before, we made use of the 20 growth clusters as the Y matrix. We created four types of data sets with varying degrees of correlation among the features, but the basic data-generation process was the same for each, and we describe it below.

We started by trying to approximate what the true distribution of nonzero coefficients in β might look like by running the RIC algorithm described in Section 2.2 with a cost of 3 bits per coefficient. The resulting 526×20 $\hat{\beta}$ matrix contained 33 nonzero coefficients, many of them in blocks of four or five down certain columns due to correlation among adjacent marker features.

The next step was to construct a true β matrix β_{sim} resembling, but not identical to, $\hat{\beta}$.²⁵ We did this based on summary statistics about $\hat{\beta}$: What fraction f of the features had any nonzero coefficients? ($f = 0.05$ for the actual $\hat{\beta}$.) Of those that did, what was the average number a of nonzero coefficients in each row? ($a = 1.3$ here.) And what were the sample mean $\mu_{\hat{\beta}}$ and standard deviation $\sigma_{\hat{\beta}}$ of the nonzero coefficient values? (Actual values were -0.12 and 0.20 , respectively.) We initialized an empty 526×20 matrix β_{sim} and, to fill it in, walked down the rows, each time flipping a coin with probability f to decide whether to give that row nonzero coefficients. We drew the number of nonzero coefficients from a Poisson distribution with rate a (capped at 20, the total number of responses) and distributed those coefficients randomly among the 20 columns. The values of each of those coefficients were drawn independently from a normal distribution with mean $\mu_{\hat{\beta}}$ and standard deviation $\sigma_{\hat{\beta}}$. When we had finished walking down the rows, we checked to make sure that the total number of nonzero coefficients in β_{sim} was within 25% of that in the original $\hat{\beta}$; if not, we started the process over.

We then constructed a simulated 100×526 X matrix X_{sim} by drawing each row from a multivariate-normal distribution with some covariance $\hat{\Sigma}^X$ based on the covariance matrix of the real X matrix. The first three variants of our synthetic data set consisted of taking $\hat{\Sigma}^X$ to be one of $\hat{\Sigma}_D^X$ (diagonal), $\hat{\Sigma}_\lambda^X$ with $\lambda = 0.5$ (half diagonal, half full), and $\hat{\Sigma}_F^X$ (full), with the subscripts as in Section 2.2.2. Figure 2 plots, respectively, $\hat{\Sigma}_D^X$, $\hat{\Sigma}_{0.5}^X$, and $\hat{\Sigma}_F^X$ as correlation matrices (i.e., each entry is scaled to fall in the range $[0, 1]$), with red indicating “high correlation” and blue, “low correlation.”

All three of the above variants involved normally distributed feature values, even though X itself consisted only of 0’s and 1’s (minor and major allele, respectively). Thus, as the fourth variant, we took X_{sim} to be the original binary-valued X matrix.

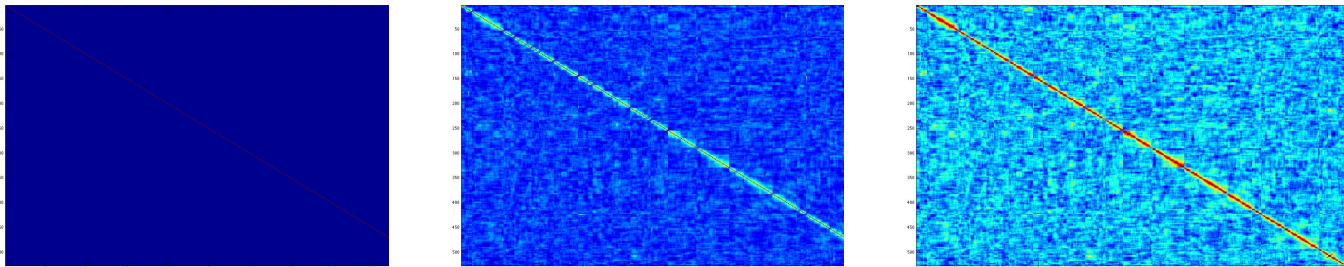


Figure 2: Correlation matrices of the 526 marker features in the Yeast Growth data set. The figures correspond to, from left to right, diagonal shrinkage, half-diagonal shrinkage, and no shrinkage. Red = correlation near 1, yellow = correlation near 0.8, light blue = correlation near 0.2, dark blue = correlation near 0. Due to the relatively high correlation among features, measures of precision and recall are slightly misleading, because which features are “truly correlated” with responses becomes unclear, as the text explains.

Finally, we estimated $\hat{\Sigma}_D$ in the manner of Section 2.2.2 with only an intercept feature in the model. With these matrices in place, we computed

$$Y_{\text{sim}} = X_{\text{sim}}\beta_{\text{sim}} + \epsilon_{\text{sim}},$$

with each row of ϵ_{sim} being independently distributed as $\mathcal{N}_h(0, \hat{\Sigma}_D)$.

Table 6 shows the results for each variant of the data set. In general, MIC performs essentially indistinguishably from the other methods. This is probably because a , the average number of responses per feature, is only 1.3 for this particular data set, implying that there isn’t substantial sharing of features across responses. There does appear to be enough sharing that MIC doesn’t perform as much worse than the other methods as in the case of the Independent synthetic data set in Table 5.

The precision of each of the methods degrades significantly as the correlation among the features increases. In fact, these numbers are somewhat misleading, because the notion of a “true coefficient” becomes fuzzy when features are correlated.

²⁵If we took β_{sim} to be equal to $\hat{\beta}$, then the RIC algorithm might have artificially high precision and recall. That’s because we’re defining the “true features” for the synthetic data set as those that the RIC algorithm returned on the real data set. Since we designed the synthetic data set to imitate the real one, we might expect RIC to return similar patterns of coefficients in both cases. Generating a new matrix to serve as β_{sim} also allows for randomization over multiple instances of this synthetic data set. Of course, one downside of this simulation approach is that correlational structure from the original problem is lost.

We generated the data by imposing a particular β_{sim} matrix, with nonzero coefficients in certain locations. For instance, for response 3, we might have given feature 186 a coefficient of 0.2, causing a correlation between that feature and that response. However, because feature 186 is correlated with, say, features 185 and 187, features 185 and 187 will also be correlated with response 3, so we end up selecting them as well. This isn't necessarily wrong to do, because the data really do show a correlation; the arbitrariness of our choice of zeros and nonzeros in β_{sim} is at fault for the large apparent rate of "false positives."

Each row of X distributed $\mathcal{N}_m(0, \widehat{\Sigma}_D^X)$ —no correlation among features.							
Method	Truth	Bonf, $\alpha=0.1$	Indep,cpc=0.1	Bonf-MIC	BH, $\alpha=0.06$	BHMat, $\alpha=0.085$	BH-MIC
Train Err	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
Test Err	0.02 ± 0.00	0.03 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
Coeff Prec	1.00 ± 0.00	0.92 ± 0.01	0.90 ± 0.01	0.92 ± 0.01	0.91 ± 0.01	0.91 ± 0.01	0.92 ± 0.01
Coeff Rec	1.00 ± 0.00	0.80 ± 0.01	0.81 ± 0.01	0.79 ± 0.01	0.81 ± 0.01	0.81 ± 0.01	0.80 ± 0.01
Each row of X distributed $\mathcal{N}_m(0, \widehat{\Sigma}_{0.5}^X)$ —some correlation among features.							
Method	Truth	Bonf, $\alpha=0.055$	Indep,cpc=1	Bonf-MIC	BH, $\alpha=0.03$	BHMat, $\alpha=0.045$	BH-MIC
Train Err	0.02 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
Test Err	0.02 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.02 ± 0.00
Coeff Prec	1.00 ± 0.00	0.78 ± 0.01	0.76 ± 0.01	0.78 ± 0.01	0.75 ± 0.01	0.75 ± 0.01	0.75 ± 0.01
Coeff Rec	1.00 ± 0.00	0.79 ± 0.01	0.79 ± 0.01	0.80 ± 0.01	0.79 ± 0.01	0.79 ± 0.01	0.80 ± 0.01
Each row of X distributed $\mathcal{N}_m(0, \widehat{\Sigma}_F^X)$ —lots of correlation among features.							
Method	Truth	Bonf, $\alpha=0.19$	Indep,cpc=0.1	Bonf-MIC	BH, $\alpha=0.07$	BHMat, $\alpha=0.06$	BH-MIC
Train Err	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
Test Err	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
Coeff Prec	1.00 ± 0.00	0.19 ± 0.01	0.20 ± 0.01	0.19 ± 0.01	0.15 ± 0.01	0.15 ± 0.00	0.15 ± 0.01
Coeff Rec	1.00 ± 0.00	0.84 ± 0.01	0.84 ± 0.01	0.84 ± 0.01	0.86 ± 0.01	0.85 ± 0.01	0.85 ± 0.01
Real, original X matrix.							
Method	Truth	Bonf, $\alpha=0.17$	Indep,cpc=0.1	Bonf-MIC	BH, $\alpha=0.06$	BHMat, $\alpha=0.055$	BH-MIC
Train Err	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
Test Err	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
Coeff Prec	1.00 ± 0.00	0.15 ± 0.00	0.15 ± 0.00	0.15 ± 0.00	0.13 ± 0.00	0.13 ± 0.00	0.13 ± 0.00
Coeff Rec	1.00 ± 0.00	0.82 ± 0.01	0.81 ± 0.01	0.83 ± 0.01	0.84 ± 0.01	0.84 ± 0.01	0.84 ± 0.01

Table 6: Test-set accuracy, precision, and recall of MIC and other methods on 25 instances of the synthetic data sets generated as described in Section 4.3.2, with $m = 526$ features.

5 Conclusion

The MDL principle provides a natural framework in which to design penalized-regression criteria for feature selection. In the case of a single response, one example of this is RIC, which can be characterized by an information-theoretic penalty of $2 \lg m$ bits per feature when selecting from among m total features. We proposed an extension of this criterion, called MIC, for the case of multiple responses. By efficiently coding the locations of feature-response pairs for features associated with multiple responses, MIC allows for sharing of information across responses during feature selection. The method is competitive with, and sometimes outperforms, existing multitask learning algorithms in terms of prediction accuracy, while achieving generally sparser, more interpretable models.

MDL can also be viewed in the domain of hypothesis testing as a way of correcting against alpha inflation in multiple tests. We explained how MDL regression applied to each feature separately can be interpreted along the lines of standard Bonferroni and Benjamini-Hochberg feature-selection procedures. Again using the MIC approach, we extended this to hypothesis testing with multiple responses, allowing for greater power in selecting true coefficients when features are significantly shared across responses.

A MDL and Hypothesis Testing

Information theory describes an isomorphism between probabilities and code lengths in which the idealized code length of some symbol x is $-\lg P(x)$ [Gru05, p. 28]. There is a similar, though only approximate, relationship between MDL and the process of *statistical hypothesis testing*: namely, that MDL will tend to introduce an extra parameter into a model in roughly the same cases that a hypothesis test would reject the null hypothesis that the parameter is zero. This is perhaps unsurprising, because MDL is a tool for model selection, and hypothesis testing is about rejecting models that fit the data poorly.

A.1 Generalized Likelihood Ratio Tests

Given two models \mathcal{M}_0 and \mathcal{M}_1 and data \mathcal{D} , statisticians define Λ to be the ratio of their likelihoods:

$$\Lambda := \frac{P(\mathcal{D} | \mathcal{M}_0)}{P(\mathcal{D} | \mathcal{M}_1)}. \quad (31)$$

In many cases, these models correspond to the same probability density function with different parameter settings. For instance, in regression of a single response y on a single feature x (with no intercept for simplicity):

$$y = \beta x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \quad (32)$$

\mathcal{M}_0 might specify that, given x , y is distributed $\mathcal{N}(0, \sigma^2)$, i.e., that $\beta = 0$, while \mathcal{M}_1 might say that given x , y is distributed $\mathcal{N}(\hat{\beta}x, \sigma^2)$, where $\hat{\beta}$ is the maximum-likelihood estimate for β derived from the training data. Note that \mathcal{M}_1 here depends on the observed data \mathcal{D} . Some authors prefer to speak of fixed probability distributions \mathcal{M}_θ with parameters θ ranging over different parameter spaces Θ_0 and Θ_1 [LM05, p. 463], in which case

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} P(\mathcal{D} | \mathcal{M}_\theta)}{\sup_{\theta \in \Theta_1} P(\mathcal{D} | \mathcal{M}_\theta)}.$$

In the regression example, $\Theta_0 = \{0\}$ while $\Theta_1 = \mathbb{R}$, the entire real line.

In fact, it will be convenient to consider $-\lg \Lambda$, which is large when Λ is small. In some sense, $-\lg \Lambda$ measures the “badness of fit” of \mathcal{M}_0 relative to \mathcal{M}_1 . Over different data sets \mathcal{D} , $-\lg \Lambda$ will take on different values, according to some probability distribution. If \mathcal{M}_0 is true, $-\ln \Lambda$ is unlikely to be very large, so we can define a threshold T_α such that

$$P(-\lg \Lambda > T_\alpha | \mathcal{M}_0) = \alpha \quad (33)$$

and reject \mathcal{M}_0 whenever $-\ln \Lambda$ exceeds this threshold. This is known as a *generalized likelihood ratio test* (GLRT), of which many standard statistical hypothesis tests are examples, including the t -test on the significance of a regression coefficient [LM05, p. 685].

A.2 Model Selection

Consider the approach that MDL would take. It “rejects” \mathcal{M}_0 in favor of \mathcal{M}_1 just in the case that the description length associated with \mathcal{M}_1 is shorter:

$$\ell(\mathcal{M}_1) + \ell(\mathcal{D} | \mathcal{M}_1) < \ell(\mathcal{M}_0) + \ell(\mathcal{D} | \mathcal{M}_0), \quad (34)$$

where $\ell(\cdot)$ stands for description length. By the discussion in section 1.2.1, $\ell(\mathcal{D} | \mathcal{M}_1) = -\lg P(\mathcal{D} | \mathcal{M}_1)$, so that (34) becomes

$$\ell(\mathcal{M}_1) - \ell(\mathcal{M}_0) < \lg P(\mathcal{D} | \mathcal{M}_1) - \lg P(\mathcal{D} | \mathcal{M}_0),$$

or

$$-\lg \Lambda > \ell(\mathcal{M}_1) - \ell(\mathcal{M}_0). \quad (35)$$

If \mathcal{M}_1 is more complicated than \mathcal{M}_0 , then $\ell(\mathcal{M}_1) - \ell(\mathcal{M}_0)$ will be positive, and there will be some α at which $\ell(\mathcal{M}_1) - \ell(\mathcal{M}_0) = T_\alpha$ as defined in (33). Thus, selecting between \mathcal{M}_0 and \mathcal{M}_1 is equivalent to doing a likelihood ratio test at the implicit significance level α determined by $\ell(\mathcal{M}_1) - \ell(\mathcal{M}_0)$.

A.3 Example: Single Regression Coefficient

While MDL introduces the quantity $-\lg \Lambda$, it is more common in statistics to deal instead with $-2 \ln \Lambda = -(2 \ln 2) \lg \Lambda$ because of the following result: If \mathcal{M}_0 and \mathcal{M}_1 both belong to the same type of probability distribution satisfying certain smoothness conditions, and if \mathcal{M}_0 has parameter space Θ_0 with dimensionality $\dim(\Theta_0)$ while \mathcal{M}_1 has parameter space Θ_1 with dimensionality $\dim(\Theta_1) > \dim(\Theta_0)$, then under \mathcal{M}_1 , $-(2 \ln 2) \lg \Lambda$ is asymptotically chi-square with degrees of freedom equal to $\dim(\Theta_1) - \dim(\Theta_0)$ [Ric95, sec. 9.5]. In particular, if the probability distribution is normal (as in our regression model), then the chi-square distribution will be not just asymptotic but exact.

In (32), where we have a single regression coefficient, $\dim(\Theta_1) - \dim(\Theta_0) = 1$, so that under \mathcal{M}_1 , $-(2 \ln 2) \lg \Lambda$ has a chi-square distribution with 1 degree of freedom, which is the same as the distribution of the square of a standard normal random variable $Z \sim \mathcal{N}(0, 1)$. Thus, we can rewrite (33) as

$$\begin{aligned} P(-(2 \ln 2) \lg \Lambda > (2 \ln 2) T_\alpha \mid \mathcal{M}_0) = \alpha &\iff P(Z^2 > (2 \ln 2) T_\alpha) = \alpha \\ &\iff P(Z < -\sqrt{(2 \ln 2) T_\alpha}) = \frac{\alpha}{2} \\ &\iff \Phi(-\sqrt{(2 \ln 2) T_\alpha}) = \frac{\alpha}{2}, \end{aligned} \tag{36}$$

where Φ is the cumulative distribution function of the standard-normal distribution. Using the approximation $\Phi(-x) \approx \frac{1}{4} e^{-\frac{x^2}{2}}$ for relatively large $x > 0$,²⁶ (36) becomes

$$\frac{1}{4} \exp\left(-\frac{(2 \ln 2) T_\alpha}{2}\right) = \frac{\alpha}{2} \iff T_\alpha = -\frac{\ln(2\alpha)}{\ln 2} = -\lg \alpha - 1. \tag{38}$$

A.4 Bonferroni Criterion

The Bonferroni criterion in p-value space says “reject H_0 when the p-value is less than $\frac{\alpha}{m}$.” By (33), we can express this in log-likelihood space by saying “reject H_0 when $-\lg \Lambda > T_{\frac{\alpha}{m}}$,” where, by (38),

$$T_{\frac{\alpha}{m}} = \lg m - \lg \alpha - 1. \tag{39}$$

Note the similarity to the decision criterion (28).

A.5 Closeness of the Approximation

It’s worth reflecting on why these correspondences are only approximate. Indeed, if code length is just negative log probability, and a p-value p is a probability, then, say, the Bonferroni rule of rejecting when $p < \frac{\alpha}{m}$ is equivalent to rejecting when $-\lg p > \lg m - \lg \alpha$, which looks basically the same as (39). However, Λ , which is used by MDL, is not exactly equal to p ; the former is a comparison of a probability density function at two points, while the latter is an area under the probability density function in extreme regions. Still, the two are often quite close in practice.

How close? Suppose we observe some value Λ_* for Λ . The associated p-value is

$$p = P(\Lambda < \Lambda_* \mid \mathcal{M}_0) = P(-2 \ln \Lambda > -2 \ln \Lambda_* \mid \mathcal{M}_0).$$

²⁶This follows from approximating

$$\Phi(-x) = \int_{-\infty}^{-x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

by the value of the integrand at $t = -x$, assuming $\frac{1}{\sqrt{2\pi}} \approx \frac{1}{4}$.

Another way to see the approximation is as follows. [P6145, pp. 63-67] proved that for $x \geq 0$,

$$0.5 - \Phi(-x) \approx \frac{1}{2} \sqrt{1 - \exp\left(-\frac{2x^2}{\pi}\right)}. \tag{37}$$

If x is large enough that $\Phi(-x)^2$ is negligible compared with $\Phi(-x)$, then we can multiply by 2 and square both sides of (37) to give

$$1 - 4\Phi(-x) \approx 1 - \exp\left(-\frac{2x^2}{\pi}\right) \iff \Phi(-x) \approx \frac{1}{4} \exp\left(-\frac{2x^2}{\pi}\right)$$

Assuming $\pi \approx 4$ gives the result.

Note that $\Phi(-x) \approx \frac{1}{x\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ tends to give a tighter approximation, especially for x bigger than ~ 2 , but it’s less mathematically convenient here.

I credit [Sti, p. 17] with inspiring the approach of using an approximation to Φ to compare hypothesis testing and coding costs.

If $-2 \ln \Lambda \sim \chi_{(k)}^2$ under \mathcal{M}_0 ,

$$p = 1 - F_{\chi_{(k)}^2}(-2 \ln \Lambda_*), \tag{40}$$

where F stands for the cumulative distribution function. Figure 3 compares Λ_* to the associated p-value for the case of $k = 1$ degree of freedom; a straight, 45-degree line would be a perfect match, but the approximation is generally correct to within a factor of 2 or 3.

We can use this curve to compute the α implied by a more complicated model \mathcal{M}_1 in (35). Letting $\Delta c := \ell(\mathcal{M}_1) - \ell(\mathcal{M}_0)$, (35) and (40) give

$$\alpha = 1 - F_{\chi_{(k)}^2}((2 \ln 2)\Delta c).$$

Table 7 shows examples for $k = 1$ degree of freedom. $\alpha = 0.05$ is achieved at $\Delta c = 2.77$ bits.²⁷

Δc (bits)	1	2	3	4
Implied α	0.24	0.1	0.04	0.02

Table 7: Example implied α values against the increase in coding cost Δc of the more complicated model, for $k = 1$ degree of freedom.

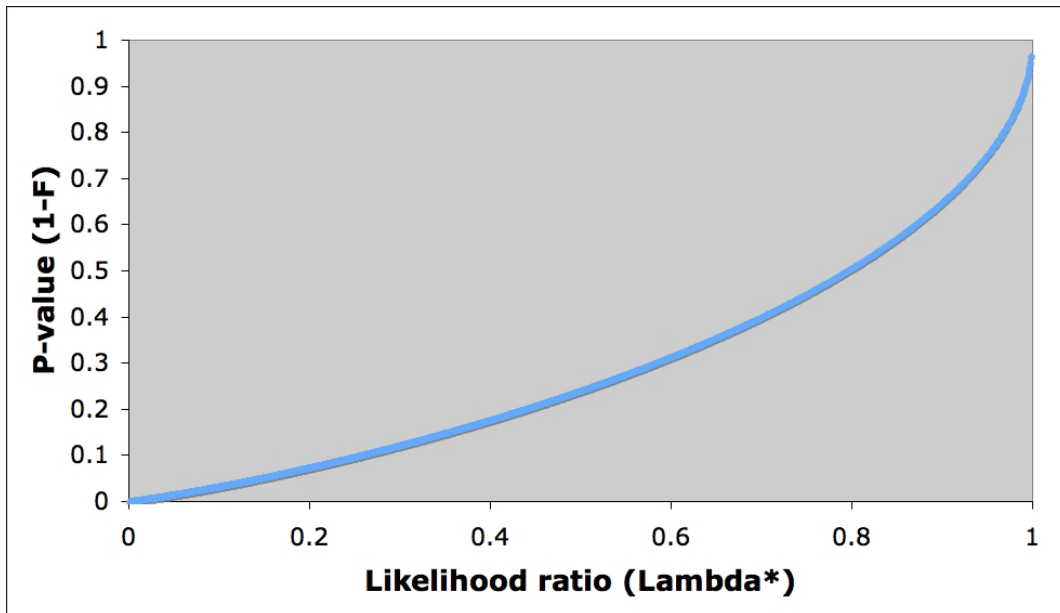


Figure 3: The actual p-value $P(\Lambda < \Lambda_* | \mathcal{M}_0)$ vs. Λ_* for $k = 1$ degree of freedom.

A.6 BH-Style Penalty

Suppose we regress y on each of m features separately. We would obtain m p-values, to which we could apply the BH step-up procedure (25). Alternatively, we could evaluate, for each feature j , a negative log-likelihood ratio:

$$-\lg \Lambda_j = -\lg P(Y | \mathcal{M}_0) - (-\lg P(Y | \mathcal{M}_{1j})),$$

where \mathcal{M}_{1j} denotes the alternative hypothesis in which feature j has a nonzero, maximum-likelihood coefficient.

In the same way that we obtained (39), we can derive a BH-type rejection level of $\frac{j\alpha}{m}$:

$$T_{\frac{j\alpha}{m}} = \lg m - \lg j - \lg \alpha - 1.$$

²⁷[BL05, p. 582] present a very similar discussion, only with log base e and with their λ equal to $\frac{\Delta c}{2}$. The numerical values given are equivalent.

We can now restate the BH procedure as follows: Put the $-\lg \Lambda_j$ in decreasing order, and let $-\lg \Lambda_{(j)}$ be the j^{th} biggest. Reject $H_{(1)}, \dots, H_{(q)}$ such that

$$q = \max \{j : -\lg \Lambda_{(j)} \geq \lg m - \lg j - \lg \alpha - 1\}. \quad (41)$$

We could rephrase (41) as choosing the q that maximizes

$$\sum_{j=1}^q (-\lg \Lambda_{(j)} - \lg m + \lg j + \lg \alpha + 1)$$

or that minimizes

$$-\sum_{j=1}^q (-\lg \Lambda_{(j)} - \lg m + \lg j + \lg \alpha + 1) = \lg \frac{m^q}{q!} - q(\lg \alpha + 1) + \sum_{j=1}^q (\lg P(Y | \mathcal{M}_0) - \lg P(Y | \mathcal{M}_{1(j)})). \quad (42)$$

Since $\lg P(Y | \mathcal{M}_0)$ is a constant with respect to q , we can subtract it m times from (42) to give

$$\lg \frac{m^q}{q!} - q(\lg \alpha + 1) + \sum_{j=1}^q -\lg P(Y | \mathcal{M}_{1(j)}) + \sum_{j=q+1}^m -\lg P(Y | \mathcal{M}_0) = \lg \frac{m^q}{q!} - q(\lg \alpha + 1) + \mathcal{D}^m(Y | \hat{\beta}_q) \quad (43)$$

with $\mathcal{D}^m(Y | \hat{\beta}_q)$ as in (26), in which $\hat{\beta}_q$ is the model containing the q features with highest negative log-likelihood ratio values. Now, if $m \gg q$, then $m^q \approx \frac{m!}{(m-q)!}$, and (43) becomes

$$\lg \frac{m!}{q!(m-q)!} - q(\lg \alpha + 1) + \mathcal{D}^m(Y | \hat{\beta}_q) = \lg \binom{m}{q} - q(\lg \alpha + 1) + \mathcal{D}^m(Y | \hat{\beta}_q), \quad (44)$$

in very close analogy to (29).²⁸

Acknowledgements

The MIC approach to regression for prediction (i.e., the content of Section 2) was originated by Jing Zhou, Dean Foster, and Lyle Ungar, who drafted a paper outlining the theory and some preliminary experiments. Lyle suggested that I continue this work as a project for summer 2008, focusing on hypothesis testing (Section 4). With continued guidance from Lyle and Dean, I kept working on the research into the fall and spring as part of this thesis. During spring 2009, Paramveer Dhillon, Lyle, and I submitted an accepted paper to ECML PKDD '09 highlighting the experimental results in Section 2. The results in Section 4 are original to this thesis.

In addition to the names above, I thank Robert Stine and Phil Everson for advice on statistical theory; Qiao Liu for conversations on MIC with Lyle; Dana Pe'er and her laboratory for providing the Yeast data set; Adam Ertel and Ted Sandler for making accessible the Breast Cancer data set; Jeff Knerr for computing assistance; Tia Newhall, Rich Wicentowski, and Doug Turnbull for guidance on writing this thesis; Santosh S. Venkatesh, my Honors thesis examiner, for several helpful comments and corrections; and two anonymous reviewers of the ECML PKDD '09 draft.

References

- [AEP08] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Mach. Learn.*, 73(3):243–272, 2008.
- [Aka73] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csàki, editors, *2nd International Symposium on Information Theory*, pages 261–281, Budapest, 1973. Akad. Kiado.
- [AM05] B. Abraham and G. Merola. Dimensionality reduction approach to multivariate prediction. *Computational Statistics and Data Analysis*, 48(1):5–16, 2005.
- [AZ05] R.K. Ando and T. Zhang. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. *The Journal of Machine Learning Research*, 6:1817–1853, 2005.

²⁸The point that MDL can reproduce BH-style step-up procedures has been noted elsewhere (e.g., [Sti, p. 19]) in slightly different words.

- [BF97] L. Breiman and J.H. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–54, 1997.
- [BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300, 1995.
- [Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.
- [BL97] A.L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.
- [BL99] Y. Benjamini and W. Liu. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference*, 82(1-2):163–170, 1999.
- [BL05] Yoav Benjamini and Moshe Leshno. *Data mining and knowledge discovery handbook*, chapter 25. Springer, 2005.
- [BVF98] PJ Brown, M. Vannucci, and T. Fearn. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 627–641, 1998.
- [BVF02] PJ Brown, M. Vannucci, and T. Fearn. Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 519–536, 2002.
- [BY98] A.R. Barron and J.B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.
- [BY01] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.
- [Car97] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.
- [CT06] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley-Interscience New York, 2006.
- [DJ94] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.
- [Eli75] P. Elias. Universal codeword sets and representations of the integers. *Information Theory, IEEE Transactions on*, 21(2):194–203, 1975.
- [FG94] D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. *Annals of Statistics*, 22:1947–1975, 1994.
- [Fri08] J. Friedman. *Fast Sparse Regression and Classification*. 2008.
- [FS99] D.P. Foster and R.A. Stine. Local asymptotic coding. *IEEE Transactions on Information Theory*, 45:1289–1293, 1999.
- [FS04] D.P. Foster and R.A. Stine. Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy. *Journal of the American Statistical Association*, 2004.
- [GF00] E.I. George and D.P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.
- [Gru05] P. Grunwald. *A Tutorial Introduction to the Minimum Description Length Principle*, chapter 1-2. MIT Press, April 2005.
- [Hoc88] Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance, 1988.
- [Hol79] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979.
- [Hom88] G. Hommel. A stagewise rejective multiple test procedure based on a modified Bonferroni test, 1988.

- [HR78] D. Harrison and D.L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.
- [HY99] M. Hansen and B. Yu. Bridging AIC and BIC: An MDL model selection criterion. In *Proceedings of the IT Workshop on Detection, Estimation, Classification and Imaging*, page 63, 1999.
- [KCY⁺06] C M Kendziorski, M Chen, M Yuan, H Lan, and A D Attie. Statistical methods for expression quantitative trait loci (eqtl) mapping. *Biometrics*, 62:19–27, March 2006. PMID: 16542225.
- [KNLG03] CM Kendziorski, MA Newton, H. Lan, and MN Gould. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in medicine*, 22(24):3899, 2003.
- [KR] Wei-Chun Kao and Alexander Rakhlin. Transfer learning toolkit. Downloaded Jan. 2009. <http://multitask.cs.berkeley.edu>.
- [KW06] C. Kendziorski and P. Wang. A review of statistical methods for expression quantitative trait loci mapping. *Mammalian Genome*, 17(6):509–517, 2006.
- [KWR⁺01] J. Khan, J.S. Wei, M. Ringnér, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7:673–679, 2001.
- [LCCP09] Oren Litvin, Helen C Causton, Bo-Juen Chen, and Dana Pe’er. Special feature: Modularity and interactions in the genetics of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, February 2009. PMID: 19223586.
- [LM05] Richard J. Larsen and Morris L. Marx. *An Introduction to Mathematical Statistics and Its Applications*. Prentice Hall, 4 edition, December 2005.
- [LPFU08] Dongyu Lin, Emily Pitler, Dean P. Foster, and Lyle H. Ungar. In defense of ℓ_0 . In *Workshop on Feature Selection at International Conference on Machine Learning (ICML 2008)*, 2008.
- [LW04] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.
- [Mal73] CL Mallows. Some Comments on C_p . *Technometrics*, 15:661–675, 1973.
- [Nat95] BK Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24:227, 1995.
- [Ng04] A.Y. Ng. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In *Proceedings of the 21st International Conference on Machine Learning*. ACM New York, NY, USA, 2004.
- [OTJ06] Guillaume Obozinski, Ben Taskar, and Michael I. Jordan. Multi-task feature selection. In *The Workshop of Structural Knowledge Transfer for Machine Learning in the 23rd International Conference on Machine Learning (ICML 2006)*, 2006.
- [OTJ09] Guillaume Obozinski, Ben Taskar, and Michael I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 2009.
- [Pó145] G. Pólya. Remarks on computing the probability integral in one and two dimensions. In *Proceedings of 1st Berkeley Symposium on Mathematics Statistics and Probabilities*, pages 63–78, 1945.
- [Ric95] J.A. Rice. *Mathematical statistics and data analysis*. Duxbury press Belmont, CA, 1995.
- [Ris78] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.
- [Ris83] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11(2):416–431, 1983.
- [Ris86] J. Rissanen. Stochastic complexity and modeling. *Ann. Statist.*, 14(3):1080–1100, September 1986.
- [Ris89] J. Rissanen. *Stochastic complexity in statistical inquiry theory*. World Scientific Publishing Co., Inc. River Edge, NJ, USA, 1989.

- [Ris99] J. Rissanen. Hypothesis selection and testing by the mdl principle. *The Computer Journal*, 42:260–269, April 1999.
- [RNK06] R. Raina, A.Y. Ng, and D. Koller. Constructing informative priors using transfer learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 713–720. ACM New York, NY, USA, 2006.
- [RYB03] A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures, 2003.
- [Sch78] Gideon Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- [Sha95] JP Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584, 1995.
- [Sim86] RJ Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.
- [SMD⁺03] Eric E. Schadt, Stephanie A. Monks, Thomas A. Drake, Aldons J. Luskis, Nam Che, Veronica Colinayo, Thomas G. Ruff, Stephen B. Milligan, John R. Lamb, Guy Cavet, Peter S. Linsley, Mao Mao, Roland B. Stoughton, and Stephen H. Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422:297–302, March 2003.
- [SORS] Juliane Schaefer, Rainer Opgen-Rhein, and Korbinian Strimmer. corpcor: Efficient estimation of covariance and (partial) correlation. <http://cran.r-project.org/web/packages/corpcor/>. Software downloaded July 2008.
- [ST05] T. Simila and J. Tikka. Multiresponse sparse regression with application to multidimensional scaling. *Lecture notes in computer science*, 3697:97, 2005.
- [ST07] T. Similä and J. Tikka. Input selection and shrinkage in multiresponse linear regression. *Computational Statistics and Data Analysis*, 52(1):406–422, 2007.
- [Sti] Robert A. Stine. Information theory and model selection. <http://www-stat.wharton.upenn.edu/~stine/research/select.info2.pdf>.
- [Sti04] Robert A. Stine. Model selection using information theory and the mdl principle. *Sociological Methods Research*, 33(2):230–260, November 2004. In-text page numbers are based on a preprint at <http://www-stat.wharton.upenn.edu/~stine/research/smr.pdf>.
- [Sto03] J.D. Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, pages 2013–2035, 2003.
- [Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [TVW05] B.A. Turlach, W.N. Venables, and S.J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.
- [vtVDvdV⁺02] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, January 2002.
- [Wal05] C.S. Wallace. *Statistical and inductive inference by minimum message length*. Springer, 2005.
- [YB99] D. Yekutieli and Y. Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1-2):171–196, 1999.