Chen and Braga-Neto *EURASIP Journal on Bioinformatics and Systems Biology* (2016) 2016:1
DOI 10.1186/s13637-015-0035-4

EURASIP Journal on Bioinformatics and Systems Biology
a SpringerOpen Journal

RESEARCH Open Access

Bayesian estimation of the discrete coefficient of determination

Ting Chen[1] and Ulisses M. Braga-Neto[2*]

Abstract

The discrete coefficient of determination (CoD) measures the nonlinear interaction between discrete predictor and target variables and has had far-reaching applications in Genomic Signal Processing. Previous work has addressed the inference of the discrete CoD using classical parametric and nonparametric approaches. In this paper, we introduce a Bayesian framework for the inference of the discrete CoD. We derive analytically the optimal minimum mean-square error (MMSE) CoD estimator, as well as a CoD estimator based on the Optimal Bayesian Predictor (OBP). For the latter estimator, exact expressions for its bias, variance, and root-mean-square (RMS) are given. The accuracy of both Bayesian CoD estimators with non-informative and informative priors, under fixed or random parameters, is studied via analytical and numerical approaches. We also demonstrate the application of the proposed Bayesian approach in the inference of gene regulatory networks, using gene-expression data from a previously published study on metastatic melanoma.

Keywords: Discrete coefficient of determination, Bayesian inference, Gene regulatory network inference

1 Introduction

DNA regulatory circuits can be often described by networks of Boolean logical gates updated and observed at discrete time intervals [1–6]. In a stochastic setting, the degree of association between Boolean predictors and targets can be quantified by means of the discrete coefficient of determination (CoD) [7]. As such, the CoD is a function of the joint probability of target and predictor variables, which, however, is usually unknown in practice. Hence, this requires the inference of the discrete CoD- given sample data. A larger sample-based CoD value indicates a tighter regulation between target and predictors.

The concept of CoD has far-reaching applications in genomics. The CoD was perhaps the first predictive paradigm utilized in the context of microarray data, the goal being to provide a measure of nonlinear interaction among genes [7]. The CoD has been used in the reconstruction or inference of gene regulatory networks using gene expression data quantized into discrete levels [8–11]. It has also been used in the definition of the intrinsically-multivariate prediction (IMP) criterion for the characterization of canalizing genes [12, 13]. In [14–16], we studied the inferential theory of the discrete CoD in a classical framework, by means of nonparametric and parametric maximum-likelihood estimation (MLE) approaches.

Classical parametric and nonparametric approaches to CoD estimation have been investigated in [14, 15]. In the present paper, we introduce a fully Bayesian approach to the inference of the discrete CoD, based on a parameterized family of target-predictor distributions. Given the priors, the probability model and sample data, we obtain the posterior distributions of the parameters, which can then be used to obtain the optimal predictors and prediction error estimators for the given problem. Such a Bayesian approach for prediction error estimation was first introduced in [17, 18], in a classification context.

Part of the work presented here appeared in [19], which introduced the *minimum mean-square error (MMSE) Bayesian CoD estimator*. In the present paper, we provide an exact representation of the analytical expressions of this estimator, and in addition, introduce the *optimal Bayesian predictor (OBP) CoD estimator*, which is based on an optimal predictor with the minimum expected true error with respect to the posterior distributions of the parameters [20, 21]. We derive exact formulas for the bias,

*Correspondence: ulisses@ece.tamu.edu
[2]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA
Full list of author information is available at the end of the article

variance, and root-mean-square (RMS) error of the OBP CoD estimator. The accuracy of both Bayesian CoD estimators is compared against that of several nonparametric CoD estimators by numerical simulations. The results indicate that the Bayesian MMSE CoD estimator is the best one when averaged over all distributions and samples, whereas the simpler OBP CoD estimator, though suboptimal in the MMSE sense, can be more accurate than the MMSE CoD estimator, in a frequentist sense, under low-variance informative priors around fixed parameters corresponding to a fixed distribution between target and predictors. It is also unsurprisingly found that priors with higher densities around true fixed distributions produce more accurate Bayesian estimators in a frequentist sense.

This paper is organized as follows. In Section 2, we introduce the discrete model for prediction and present the coefficient of determination in this model. In Section 3, we develop a Bayesian framework of the inference of the discrete CoD, define two Bayesian CoD estimators, one in the sense of minimum mean-square error (MMSE), and the other based on the optimal Bayesian classifier, and derive the analytical expressions for both Bayesian CoD estimators. In Section 4, we first present an exact formulation of accuracy metrics for the OBP CoD estimator. Afterwards, we discuss the accuracy of both Bayesian CoD estimators when averaged over all distributions and samples as well as under fixed distributions under varying priors, and their comparison with the nonparametric CoD estimators. Section 6 describes an approach to the inference of gene regulatory networks using the proposed Bayesian CoD estimators and illustrates the approach with gene expression data from a previously published study on metastatic melanoma. Finally, Section 7 presents concluding remarks.

2 The discrete coefficient of determination

The CoD, which was originally defined in classical regression analysis, gives the relative decrease in unexplained variability when entering a variable X into the regression of the dependent variable Y, in comparison with the total unexplained variability when entering no variables. Dougherty and collaborators extended the concept of CoD to discrete random variables [7]. Given a specified error criterion, such as the mean-square error or the mean-absolute error, the CoD was defined in [7] as

$$\mathrm{CoD} = \frac{\varepsilon_0 - \varepsilon}{\varepsilon_0}, \tag{1}$$

where ε_0 is the minimum error of predicting Y by a constant (i.e., in the absence of observations) and ε is the minimum error of predicting Y based on the observation of X. Since $\varepsilon \leq \varepsilon_0$ (all sensible error criteria satisfy this property), the CoD ranges from 0 to 1. The closer it is to one, the closer ε is to zero and the tighter the association between predictor and target variables, whereas the closer it is to zero, the closer ε is to ε_0 and the weaker the association is. By convention, $\mathrm{CoD} = 0$ when $\varepsilon_0 = 0$. The CoD is a function only of the distribution of (X, Y); in particular, it is not a function of sample data. This definition of the CoD reduces gracefully to the classical one in the case when (X, Y) is jointly Gaussian [7].

We consider in this paper the case where $\mathbf{X} = (X_1, X_2, \ldots, X_d) \in \{0,1\}^d$ is a binary vector of predicting variables and $Y \in \{0,1\}$ is a binary target random variable. For example, $\mathbf{X}$ and Y may consist of the active/inactive expression state of various genes. The probability distribution of the pair $(\mathbf{X}, Y)$ is specified by the probability $c = P(Y = 0)$, and the probabilities $p_i = P(\mathbf{X} = \mathbf{x}^i \mid Y = 0)$ and $q_i = P(\mathbf{X} = \mathbf{x}^i \mid Y = 1)$, for $i = 1, \ldots, b$, with $\sum_{i=1}^{2^d} p_i = 1$ and $\sum_{i=1}^{2^d} q_i = 1$. Let $\left(\mathbf{x}^1, \ldots, \mathbf{x}^{2^d}\right)$ be an arbitrary enumeration of the possible values of the predicting vector $\mathbf{X}$. An optimal predictor of Y given $\mathbf{X}$ is well-known to be $\psi^*(\mathbf{X}) = \arg\max_k P(Y = k \mid \mathbf{X})$ [22]. The minimum error of predicting Y based on the observation of $\mathbf{X}$ is therefore

$$\begin{aligned}
\varepsilon &= P(Y \neq \psi^*(\mathbf{X})) = E\left[\min\{P(Y=1 \mid X), P(Y=0 \mid X)\}\right] \\
&= \sum_{i=1}^{2^d} \min\left\{P\left(Y=1 \mid \mathbf{X}=\mathbf{x}^i\right), P\left(Y=0 \mid \mathbf{X}=\mathbf{x}^i\right)\right\} \\
&\quad \times P\left(\mathbf{X}=\mathbf{x}^i\right) \\
&= \sum_{i=1}^{2^d} \min\left\{P\left(Y=1, \mathbf{X}=\mathbf{x}^i\right), P\left(Y=0, \mathbf{X}=\mathbf{x}^i\right)\right\} \\
&= \sum_{i=1}^{2^d} \min\left\{P\left(\mathbf{X}=\mathbf{x}^i \mid Y=1\right) P(Y=1), P(\mathbf{X}=\mathbf{x}^i \mid \right. \\
&\qquad \left. Y=0) P(Y=0)\right\} \\
&= \sum_{i=1}^{2^d} \min\left\{c\, p_i, (1-c) q_i\right\} \\
&= \sum_{i=1}^{2^d} \left(c\, p_i I_{p_i < \frac{1-c}{c} q_i} + (1-c) q_i I_{q_i \leq \frac{c}{1-c} p_i}\right),
\end{aligned} \tag{2}$$

where I_A is the indicator function, which is equal to 1 if A is satisfied and zero, otherwise. On the other hand, an optimal predictor in the absence of observations is clearly given by $\psi^* = \arg\max_k P(Y = k)$, so that the minimum error of predicting Y by a constant is given by

$$\varepsilon_0 = \min\{P(Y=0), P(Y=1)\} = \min\{c, 1-c\}. \tag{3}$$

Plugging (2) and (3) in (1) results in

$$\mathrm{CoD} = 1 - \sum_{i=1}^{2^d} \left(\frac{c}{\min\{c, 1-c\}} p_i I_{p_i < \frac{1-c}{c} q_i} + \frac{1-c}{\min\{c, 1-c\}} q_i I_{q_i \leq \frac{c}{1-c} p_i} \right). \quad (4)$$

This formula gives the relationship between the CoD and the parameters of the distribution of $(\mathbf{X}, Y)$.

3 Bayesian CoD estimators

In practice, the distributional parameters are generally unknown, and one would like to estimate the CoD from sample data. We present in this section the derivation of two Bayesian estimators for the CoD in (4). One approach is analogous to that followed by [17] in defining the Bayesian MMSE prediction error estimator, whereas the other one makes use of the *optimal Bayesian predictor* (OBP), a straightforward generalization of the optimal Bayesian classifier (OBC), introduced in [20].

We will assume that an i.i.d. sample $\mathbf{S}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ from the distribution of $(\mathbf{X}, Y)$ is available. Given $\mathbf{S}_n$, define U_i as the number of sample points such that $\mathbf{X} = \mathbf{x}^i$ and $Y = 0$, and V_i as the number of sample points such that $\mathbf{X} = \mathbf{x}^i$ and $Y = 1$, for $i = 1, \ldots, 2^d$. Note that $N_0 = \sum_{i=1}^{2^d} U_i$ and $N_1 = \sum_{i=1}^{2^d} V_i$ are the (random) sample sizes corresponding to $Y = 0$ and $Y = 1$, respectively.

Let $\mathbf{p} = (p_1, \ldots, p_{2^d})$, $\mathbf{q} = (q_1, \ldots, q_{2^d})$, and $\boldsymbol{\theta} = (c, \mathbf{p}, \mathbf{q})$, where $0 \leq c, p_i, q_i \leq 1$, and $\sum_{i=1}^{2^d} p_i = \sum_{i=1}^{2^d} q_i = 1$. As shown in the previous section, the distribution of $(\mathbf{X}, Y)$ is completely specified by the parameter vector $\boldsymbol{\theta}$. The Bayesian approach treats $\boldsymbol{\theta}$ as a random variable, the *prior distribution* of which can take advantage of a priori knowledge about the problem. We will assume that c, $\mathbf{p}$, and $\mathbf{q}$ are independent, i.e., $f(\boldsymbol{\theta}) = f(c)f(\mathbf{p})f(\mathbf{q})$. It is shown in [17] that this implies that the *posterior distribution* of $\boldsymbol{\theta}$ also factors $f(\boldsymbol{\theta} \mid \mathbf{S}_n) = f(c \mid \mathbf{S}_n) f(\mathbf{p} \mid \mathbf{S}_n) f(\mathbf{q} \mid \mathbf{S}_n)$.

In this paper, we will employ the standard choice of priors for discrete distributions, namely, the Beta and Dirichlet distributions (c.f. Appendices A and B):

$$\begin{aligned} c &\sim \mathrm{Beta}(\alpha, \beta), \\ \mathbf{p} &\sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_{2^d}), \\ \mathbf{q} &\sim \mathrm{Dirichlet}(\beta_1, \ldots, \beta_{2^d}), \end{aligned} \quad (5)$$

where the hyperparameters α, β, α_i, β_i, $i = 1, \ldots, 2^d$, are positive numbers. These distributions have bounded supports; the Beta distribution is defined over the interval $[0, 1]$, while the Dirichlet distribution is defined over the simplex of 2^d nonnegative numbers that add up to one. The shapes of the distributions are controlled by the *concentration parameters* $\Delta_c = \alpha + \beta$, $\Delta_p = \sum_{j=1}^{2^d} \alpha_j$, and $\Delta_q = \sum_{j=1}^{2^d} \beta_j$, and the *base measures* $c_0 = \alpha/\Delta_c$, $\mathbf{p}_0 = (\alpha_1/\Delta_p, \ldots, \alpha_{2^d}/\Delta_p)$, and $\mathbf{q}_0 = (\beta_1/\Delta_q, \ldots, \beta_{2^d}/\Delta_q)$. Please refer to Appendices A and B for definitions and important facts about the Beta and Dirichlet distributions, which will be needed in the sequel.

A very important property for our purposes is that the Beta and Dirichlet priors are *conjugate priors* for the discrete multinomial distribution, i.e., they have the same form as the corresponding posteriors. Given the sample data $\mathbf{S}_n$, the posterior distributions are [17, 18]:

$$\begin{aligned} c \mid \mathbf{S}_n &\sim \mathrm{Beta}(n_0 + \alpha, n_1 + \beta), \\ \mathbf{p} \mid \mathbf{S}_n &\sim \mathrm{Dirichlet}(u_1 + \alpha_1, \ldots, u_{2^d} + \alpha_{2^d}), \\ \mathbf{q} \mid \mathbf{S}_n &\sim \mathrm{Dirichlet}(v_1 + \beta_1, \ldots, v_{2^d} + \beta_{2^d}). \end{aligned} \quad (6)$$

where n_0 and n_1 are the observed sample sizes corresponding to $Y = 0$ and $Y = 1$, respectively, while u_i and v_i are the observed sample values of the random variables U_i and V_i, respectively.

3.1 Minimum mean-square error CoD estimator

Given a CoD estimator $\widehat{\mathrm{CoD}}$, consider the mean-square error

$$\mathrm{MSE} = E_{\boldsymbol{\theta}, \mathbf{S}_n} \left[|\widehat{\mathrm{CoD}} - \mathrm{CoD}|^2 \right]. \quad (7)$$

The minimum MSE solution, as is well known, is given by the expectation of the CoD according to the posterior distribution of the parameters [23]. This defines the *Bayesian MMSE CoD estimator*:

$$\widehat{\mathrm{CoD}}_{\mathrm{MMSE}} = E[\mathrm{CoD} \mid \mathbf{S}_n] = E_{\boldsymbol{\theta} \mid \mathbf{S}_n}[\mathrm{CoD}], \quad (8)$$

where the CoD is given by (4).

It is well-known that the MMSE estimator $\widehat{\mathrm{CoD}}_{\mathrm{MMSE}}$ not only displays the least root mean-square error (RMS) over the distribution of $(\boldsymbol{\theta}, \mathbf{S}_n)$, but it is also an unbiased estimator (however, for a specific model with fixed $\boldsymbol{\theta}$, $\widehat{\mathrm{CoD}}_{\mathrm{MMSE}}$ might not be unbiased or have the least RMS).

In order to derive an expression for the Bayesian MMSE CoD estimator, first note that (4) can be rewritten as

$$\mathrm{CoD} = 1 - \sum_{i=1}^{2^d} \left(p_i I_{p_i < \frac{1-c}{c} q_i} I_{c < 1/2} + \frac{c}{1-c} p_i I_{p_i < \frac{1-c}{c} q_i} I_{c \geq 1/2} + \frac{1-c}{c} q_i I_{q_i \leq \frac{c}{1-c} p_i} I_{c < 1/2} + q_i I_{q_i \leq \frac{c}{1-c} p_i} I_{c \geq 1/2} \right). \quad (9)$$

Applying (8) to (9) and using the previously mentioned fact that the posterior distribution factors allows one to write the Bayesian MMSE CoD estimator as

$$
\begin{aligned}
\widehat{\mathrm{CoD}}_{\mathrm{MMSE}} &= E_{\boldsymbol{\theta}|\mathbf{S}_n}[\mathrm{CoD}] = E_{c|\mathbf{S}_n}\left[E_{\mathbf{p}|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}[\mathrm{CoD}]\right]\right] \\
&= 1 - \sum_{i=1}^{2^d}\Bigg(E_{c|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[E_{\mathbf{p}|\mathbf{S}_n}\left[p_i I_{p_i<\frac{1-c}{c}q_i}\right] I_{c<1/2}\right]\right] \\
&\quad + E_{c|\mathbf{S}_n}\left[\frac{c}{1-c}E_{\mathbf{q}|\mathbf{S}_n}\left[E_{\mathbf{p}|\mathbf{S}_n}\left[p_i I_{p_i<\frac{1-c}{c}q_i}\right] I_{c\geq 1/2}\right]\right] \\
&\quad + E_{c|\mathbf{S}_n}\left[\frac{1-c}{c}E_{\mathbf{p}|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[q_i I_{q_i\leq\frac{c}{1-c}p_i}\right] I_{c<1/2}\right]\right] \\
&\quad + E_{c|\mathbf{S}_n}\left[E_{\mathbf{p}|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[q_i I_{q_i\leq\frac{c}{1-c}p_i}\right] I_{c\geq 1/2}\right]\right]\Bigg).
\end{aligned} \tag{10}
$$

Using (6) and the fact that the marginal distributions of a Dirichlet are Beta (c.f. Appendix B), we have that $c \mid \mathbf{S}_n \sim \mathrm{Beta}(\alpha^s, \beta^s)$, $p_i \mid \mathbf{S}_n \sim \mathrm{Beta}(\alpha_i^s, \overline{\alpha}_i^s)$, and $q_i \mid \mathbf{S}_n \sim \mathrm{Beta}(\beta_i^s, \overline{\beta}_i^s)$, where $\alpha^s = n_0 + \alpha$, $\beta^s = n_1 + \beta$, $\alpha_i^s = u_i + \alpha_i$, $\overline{\alpha}_i^s = n_0 - u_i + \Delta_p - \alpha_i$, $\beta_i^s = \nu_i + \beta_i$, and $\overline{\beta}_i^s = n_1 - \nu_i + \Delta_q - \beta_i$, for $i = 1, \ldots, 2^d$. Using the results in Appendix A and assuming that the hyperparameters are integers (if they are not, a simple adjustment to the derivation below can be made; see Appendix A), it follows that

$$
\begin{aligned}
&E_{c|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[E_{\mathbf{p}|\mathbf{S}_n}\left[p_i I_{p_i<\frac{1-c}{c}q_i}\right] I_{c<1/2}\right]\right] = E_{c|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[E_{\mathbf{p}|\mathbf{S}_n}\left[p_i I_{p_i<\frac{1-c}{c}q_i}\right] I_{q_i<\frac{c}{1-c}}\right] I_{c<1/2}\right] + \\
&E_{c|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[E_{\mathbf{p}|\mathbf{S}_n}[p_i]\right] I_{c<1/2}\right] - E_{c|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[E_{\mathbf{p}|\mathbf{S}_n}[p_i] I_{q_i<\frac{c}{1-c}}\right] I_{c<1/2}\right] \\
&= \frac{1}{\mathrm{B}(\alpha_i^s, \overline{\alpha}_i^s)} \times \Bigg\{E_{c|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[\sum_{j=0}^{\overline{\alpha}_i^s-1} r_j(\alpha_i^s+1, \overline{\alpha}_i^s)\left(\frac{1-c}{c}q_i\right)^{\alpha_i^s+j+1} I_{q_i<\frac{c}{1-c}}\right] I_{c<1/2}\right] + \\
&E_{c|\mathbf{S}_n}\left[\mathrm{B}\left(\alpha_i^s+1, \overline{\alpha}_i^s\right) I_{c<1/2}\right] - E_{c|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[\mathrm{B}\left(\alpha_i^s+1, \overline{\alpha}_i^s\right) I_{q_i<\frac{c}{1-c}}\right] I_{c<1/2}\right]\Bigg\} = \frac{1}{\mathrm{B}\left(\alpha_i^s, \overline{\alpha}_i^s\right)\mathrm{B}\left(\beta_i^s, \overline{\beta}_i^s\right)} \times \\
&\Bigg\{\sum_{j=0}^{\overline{\alpha}_i^s-1}\sum_{k=0}^{\overline{\beta}_i^s-1} r_j\left(\alpha_i^s+1, \overline{\alpha}_i^s\right) r_k\left(\alpha_i^s+\beta_i^s+j+1, \overline{\beta}_i^s\right) E_{c|\mathbf{S}_n}\left[\left(\frac{c}{1-c}\right)^{\beta_i^s+k} I_{c<1/2}\right] + \\
&\mathrm{B}\left(\alpha_i^s+1, \overline{\alpha}_i^s\right)\mathrm{B}\left(\beta_i^s, \overline{\beta}_i^s\right) E_{c|\mathbf{S}_n}\left[I_{c<1/2}\right] - \mathrm{B}\left(\alpha_i^s+1, \overline{\alpha}_i^s\right)\sum_{j=0}^{\overline{\beta}_i^s-1} r_j\left(\beta_i^s, \overline{\beta}_i^s\right) E_{c|\mathbf{S}_n}\left[\left(\frac{c}{1-c}\right)^{\beta_i^s+j} I_{c<1/2}\right]\Bigg\} \\
&= \frac{1}{2^{\alpha^s}\,\mathrm{B}(\alpha^s, \beta^s)\mathrm{B}\left(\alpha_i^s, \overline{\alpha}_i^s\right)\mathrm{B}\left(\beta_i^s, \overline{\beta}_i^s\right)} \times \\
&\times \Bigg\{\sum_{j=0}^{\overline{\alpha}_i^s-1}\sum_{k=0}^{\overline{\beta}_i^s-1}\sum_{l=0}^{\beta^s-(\beta_i^s+k+1)}\Bigg[r_j\left(\alpha_i^s+1, \overline{\alpha}_i^s\right) r_k\left(\alpha_i^s+\beta_i^s+j+1, \overline{\beta}_i^s\right) r_l\left(\alpha^s+\beta_i^s+k, \beta^s-\left(\beta_i^s+k\right)\right) \\
&\quad \times \frac{1}{2^{\beta_i^s+k+l}}\Bigg] + \mathrm{B}\left(\alpha_i^s+1, \overline{\alpha}_i^s\right)\mathrm{B}\left(\beta_i^s, \overline{\beta}_i^s\right)\sum_{j=0}^{\beta^s-1} r_j\left(\alpha^s, \beta^s\right)\frac{1}{2^j} \\
&\quad - \mathrm{B}\left(\alpha_i^s+1, \overline{\alpha}_i^s\right)\sum_{j=0}^{\overline{\beta}_i^s-1}\sum_{k=0}^{\beta^s-(\beta_i^s+j+1)} r_j\left(\beta_i^s, \overline{\beta}_i^s\right) r_k\left(\alpha^s+\beta_i^s+j, \beta^s-\left(\beta_i^s+j\right)\right)\frac{1}{2^{\beta_i^s+j+k}}\Bigg\}.
\end{aligned} \tag{11}
$$

Likewise, we have

$$
\begin{aligned}
&E_{c|\mathbf{S}_n}\left[\frac{c}{1-c}E_{\mathbf{q}|\mathbf{S}_n}\left[E_{\mathbf{p}|\mathbf{S}_n}\left[p_i I_{p_i<\frac{1-c}{c}q_i}\right] I_{c\geq 1/2}\right]\right] = \frac{1}{2^{\beta^s}\,\mathrm{B}(\alpha^s, \beta^s)\mathrm{B}\left(\alpha_i^s, \overline{\alpha}_i^s\right)\mathrm{B}\left(\beta_i^s, \overline{\beta}_i^s\right)} \\
&\times \Bigg\{\sum_{j=0}^{\overline{\alpha}_i^s-1}\sum_{k=0}^{\alpha^s-(\alpha_i^s+j+1)}\Bigg[r_j\left(\alpha_i^s+1, \overline{\alpha}_i^s\right) r_k\left(\beta^s+\alpha_i^s+j, \alpha^s-\left(\alpha_i^s+j\right)\right)\mathrm{B}\left(\alpha_i^s+\beta_i^s+j+1, \overline{\beta}_i^s\right) \\
&\quad \times \frac{1}{2^{\alpha_i^s+j+k}}\Bigg]\Bigg\},
\end{aligned} \tag{12}
$$

$$
\begin{aligned}
&E_{c|\mathbf{S}_n}\left[\frac{1-c}{c}E_{\mathbf{p}|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[q_i\, I_{q_i\le\frac{c}{1-c}p_i}\right]I_{c<1/2}\right]\right] \\
&= \frac{1}{2^{\alpha^s}\,\mathrm{B}(\alpha^s,\beta^s)\mathrm{B}\left(\alpha_i^s,\overline{\alpha}_i^s\right)\mathrm{B}\left(\beta_i^s,\overline{\beta}_i^s\right)} \\
&\times\left\{\sum_{j=0}^{\overline{\beta}_i^s-1}\sum_{k=0}^{\beta^s-(\beta_i^s+j+1)}\left[r_j\left(\beta_i^s+1,\overline{\beta}_i^s\right)r_k(\alpha^s+\beta_i^s+j,\right.\right. \\
&\quad \beta^s-(\beta_i^s+j))\,\mathrm{B}\left(\alpha_i^s+\beta_i^s+j+1,\overline{\alpha}_i^s\right) \\
&\quad\left.\left.\times\frac{1}{2^{\beta_i^s+j+k}}\right]\right\},
\end{aligned}
\tag{13}
$$

and

$$
\begin{aligned}
&E_{c|\mathbf{S}_n}\left[E_{\mathbf{p}|\mathbf{S}_n}\left[E_{\mathbf{q}|\mathbf{S}_n}\left[q_i\, I_{q_i\le\frac{c}{1-c}p_i}\right]I_{c\ge1/2}\right]\right] \\
&= \frac{1}{2^{\beta^s}\,\mathrm{B}(\alpha^s,\beta^s)\mathrm{B}\left(\alpha_i^s,\overline{\alpha}_i^s\right)\mathrm{B}\left(\beta_i^s,\overline{\beta}_i^s\right)} \\
&\times\left\{\sum_{j=0}^{\overline{\beta}_i^s-1}\sum_{k=0}^{\overline{\alpha}_i^s-1}\sum_{l=0}^{\alpha^s-(\alpha_i^s+k+1)}\left[r_j(\beta_i^s+1,\overline{\beta}_i^s)\,r_k(\alpha_i^s+\beta_i^s+j\right.\right. \\
&\quad +1,\overline{\alpha}_i^s)r_l(\beta^s+\alpha_i^s+k,\alpha^s-(\alpha_i^s+k)) \\
&\quad\left.\times\frac{1}{2^{\alpha_i^s+k+l}}\right]+\mathrm{B}(\beta_i^s+1,\overline{\beta}_i^s)\mathrm{B}(\alpha_i^s,\overline{\alpha}_i^s)\sum_{j=0}^{\alpha^s-1}r_j(\beta^s,\alpha^s)\frac{1}{2^j} \\
&\quad-\mathrm{B}(\beta_i^s+1,\overline{\beta}_i^s)\sum_{j=0}^{\overline{\alpha}_i^s-1}\sum_{k=0}^{\alpha^s-(\alpha_i^s+j+1)}r_j(\alpha_i^s,\overline{\alpha}_i^s)\,r_k(\beta^s+\alpha_i^s \\
&\quad\left.+j,\alpha^s-(\alpha_i^s+j))\frac{1}{2^{\alpha_i^s+j+k}}\right\},
\end{aligned}
\tag{14}
$$

where the Beta function $B(a,b)$ and the coefficients $r_i(a,b)$ are defined in Appendix A.

Replacing (11)–(14) into (10) produces an exact expression for computing the MMSE CoD estimator in terms of sample sizes and model hyperparameters. Notice that for the previous expressions to make sense, one must have $\alpha>\Delta_p-1$ and $\beta>\Delta_q-1$. In particular, if uniform priors are chosen for $\mathbf{p}$ or $\mathbf{q}$, then the prior for c cannot be uniform (c.f. Appendix A).

3.2 Optimal Bayesian predictor CoD estimator

In this section, we derive a second Bayesian CoD estimator, using the *optimal Bayesian predictor* (OBP), a simple extension to the Boolean prediction problem of the "optimal Bayesian classifier" (OBC) proposed in [20]. Formally, let $\varepsilon_{\boldsymbol{\theta}}[\,\psi\,]$ denote the error of a predictor ψ under parameter vector $\boldsymbol{\theta}$. The OBP predictor ψ_{OBP} minimizes the average error over the family of (posterior) distributions indexed by the parameter

$$
\psi_{\mathrm{OBP}} = \arg\min_{\psi\in\Upsilon} E_{\boldsymbol{\theta}|\mathbf{S}_n}[\varepsilon_{\boldsymbol{\theta}}[\psi]\,]\,. \tag{15}
$$

Using the results of [20] for the OBC, one can verify that the OBP for the Beta-Dirichlet model considered here is given by

$$
\psi_{\mathrm{OBP}}(\mathbf{x}^i) = \begin{cases} 1, \text{ if } \frac{n_0+\alpha}{n+\alpha+\beta}\,\frac{U_i+\alpha_i}{n_0+\Delta_p} < \frac{n_1+\beta}{n+\alpha+\beta}\,\frac{V_i+\beta_i}{n_1+\Delta_q}, \\ 0, \text{ otherwise}, \end{cases} \tag{16}
$$

for $i=1,\ldots,2^d$, with optimal prediction error

$$
\hat{\varepsilon}_{\mathrm{OBP}} = E_{\boldsymbol{\theta}|\mathbf{S}_n}[\varepsilon_{\boldsymbol{\theta}}[\psi_{\mathrm{OBP}}]\,] = \sum_{i=1}^{2^d}\min\left\{\frac{n_0+\alpha}{n+\alpha+\beta}\,\frac{U_i+\alpha_i}{n_0+\Delta_p}, \frac{n_1+\beta}{n+\alpha+\beta}\,\frac{V_i+\beta_i}{n_1+\Delta_q}\right\}. \tag{17}
$$

On the other hand, the average errors of the the constant predictors $\psi\equiv0$ and $\psi\equiv1$ are

$$
\begin{aligned}
E_{c|\mathbf{S}_n}[P(Y=1)] &= E_{c|\mathbf{S}_n}[c] = \frac{n_0+\alpha}{n+\alpha+\beta}, \\
E_{c|\mathbf{S}_n}[P(Y=0)] &= 1-E_{c|\mathbf{S}_n}[c] = \frac{n_1+\beta}{n+\alpha+\beta},
\end{aligned}
\tag{18}
$$

respectively, so that the OBP error in the absence of observations is

$$
\hat{\varepsilon}_{0,\mathrm{OBP}} = \min\left\{\frac{n_0+\alpha}{n+\alpha+\beta}, \frac{n_1+\beta}{n+\alpha+\beta}\right\}. \tag{19}
$$

We can then combine (17) and (19) to obtain the optimal Bayesian predictor (OBP) CoD estimator

$$
\begin{aligned}
\widehat{\mathrm{CoD}}_{\mathrm{OBP}} &= 1-\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{\hat{\varepsilon}_{0,\mathrm{OBP}}} \\
&= 1-\frac{1}{\min\{n_0+\alpha,n_1+\beta\}}\sum_{i=1}^{2^d}\min \\
&\quad\left\{\frac{n_0+\alpha}{n_0+\Delta_p}(U_i+\alpha_i), \frac{n_1+\beta}{n_1+\Delta_q}(V_i+\beta_i)\right\}.
\end{aligned}
\tag{20}
$$

It is easy to show that $0\le\hat{\varepsilon}_{\mathrm{OBP}}\le\hat{\varepsilon}_{0,\mathrm{OBP}}$, and thus $0\le\widehat{\mathrm{CoD}}_{\mathrm{OBP}}\le1$.

Execution time for computation of the OBP CoD estimator grows as $O(2^d)$. By comparison, the complexity for exact computation of the Bayesian MMSE CoD estimator introduced in the previous subsection is $O(n^3\times2^d)$. Neither n or d tends to be too large in Genomics applications, due to small sample sizes and the fact that the average number of predictor genes d per target gene must be small for a stable system, as remarked by S. Kauffman in [2]. However, if n and d become large, one could devise

Monte Carlo approximation methods to compute both CoD estimators.

Therefore, the OBP CoD estimator, though suboptimal, is much more efficient computationally than the MMSE CoD estimator, especially at large sample sizes. In addition, we will see in the next section that the OBP CoD can be even more accurate than the MMSE CoD estimator, in frequentist sense, under a fixed value of the parameters.

4 Performance analysis

In this section, we investigate the accuracy of the Bayesian CoD estimators proposed in the previous section. We distinguish between two types of accuracy metrics: *global* metrics concern the average performance over all samples and all distributions of $(\mathbf{X}, Y)$, weighted by the prior distribution of $\boldsymbol{\theta}$, whereas *fixed-parameter* metrics have to do with the average performance over all samples, but under a particular distribution of $(\mathbf{X}, Y)$, corresponding to a fixed value of the parameter $\boldsymbol{\theta}$. Fixed-parameter metrics thus evaluate the proposed Bayesian estimators from a purely frequentist perspective.

For a given Bayesian CoD estimator, the fixed-parameter accuracy metrics of interest are the bias

$$\text{Bias}(\boldsymbol{\theta}) = E_{\mathbf{S}_n|\boldsymbol{\theta}}\left[\widehat{\text{CoD}} - \text{CoD}\right] = E_{\mathbf{S}_n|\boldsymbol{\theta}}\left[\frac{\hat{\varepsilon}}{\hat{\varepsilon}_0}\right] - \frac{\varepsilon}{\varepsilon_0}, \tag{21}$$

the variance,

$$\text{Variance}(\boldsymbol{\theta}) = \text{Var}_{\mathbf{S}_n|\boldsymbol{\theta}}\left[\widehat{\text{CoD}}\right] = E_{\mathbf{S}_n|\boldsymbol{\theta}}\left[\frac{\hat{\varepsilon}^2}{\hat{\varepsilon}_0^2}\right] - \left(E_{\mathbf{S}_n|\boldsymbol{\theta}}\left[\frac{\hat{\varepsilon}}{\hat{\varepsilon}_0}\right]\right)^2, \tag{22}$$

and the root-mean-square (RMS) error,

$$\begin{aligned}\text{RMS}(\boldsymbol{\theta}) &= \sqrt{E_{\mathbf{S}_n|\boldsymbol{\theta}}\left[\left(\widehat{\text{CoD}} - \text{CoD}\right)^2\right]} \\ &= \sqrt{\text{Variance}(\boldsymbol{\theta}) + \text{Bias}(\boldsymbol{\theta})^2}.\end{aligned} \tag{23}$$

It becomes clear that the fixed-parameter bias, variance, and RMS of a Bayesian CoD estimator can be obtained with knowledge of the first and second moments $E_{\mathbf{S}_n|\boldsymbol{\theta}}\left[\frac{\hat{\varepsilon}}{\hat{\varepsilon}_0}\right]$ and $E_{\mathbf{S}_n|\boldsymbol{\theta}}\left[\frac{\hat{\varepsilon}^2}{\hat{\varepsilon}_0^2}\right]$.

The corresponding global accuracy metrics are obtained by taking expectation of the previous quantities with respect to the marginal (i.e., prior) distribution of $\boldsymbol{\theta}$.

As mentioned previously, the global bias of the Bayesian MMSE CoD estimator is zero and its global RMS is minimal among all CoD estimators. However, this does not imply that its fixed-parameter bias is zero or that its fixed-parameter RMS is minimum for all values of the parameter.

In what follows, we give exact expressions for the computation of $E_{\mathbf{S}_n|\boldsymbol{\theta}}\left[\frac{\hat{\varepsilon}}{\hat{\varepsilon}_0}\right]$ and $E_{\mathbf{S}_n|\boldsymbol{\theta}}\left[\frac{\hat{\varepsilon}^2}{\hat{\varepsilon}_0^2}\right]$ for the OBP CoD estimator. As argued previously, this allows the exact computation of the fixed-parameter bias, variance, and RMS of that CoD estimator. Via simple numerical integration, it is possible then to obtain the global bias, variance, and RMS. It turns out that similar expressions for the MMSE CoD estimator are much harder to obtain; the performance of that estimator are studied via a numerical approach in the next section.

All the expectations and probabilities below are with respect to $\mathbf{S}_n \mid \boldsymbol{\theta}$ (the subscript will be omitted for convenience). In the expressions below, c, p_i, and q_i, for $i = 1, \ldots, 2^d$ refer to the (deterministic) parameters in $\boldsymbol{\theta}$.

First note that

$$\begin{aligned}E\left[\frac{\hat{\varepsilon}_{\text{OBP}}}{\hat{\varepsilon}_{0,\text{OBP}}}\right] &= E\left[E\left[\frac{\hat{\varepsilon}_{\text{OBP}}}{\hat{\varepsilon}_{0,\text{OBP}}} \;\middle|\; \hat{\varepsilon}_{0,\text{OBP}}\right]\right] \\ &= \sum_{m\in L} E\left[\frac{\hat{\varepsilon}_{\text{OBP}}}{m/(n+\alpha+\beta)} \;\middle|\; M = m\right] P(M = m),\end{aligned} \tag{24}$$

where $M = (n+\alpha+\beta)\,\hat{\varepsilon}_{0,\text{OBP}} = \min(n_0+\alpha, n_1+\beta)$ and

$$L = \left\{\alpha, \alpha+1, \ldots, \alpha + \left\lfloor \frac{n+\beta-\alpha}{2} \right\rfloor\right\} \bigcup \left\{\beta, \beta+1, \ldots, \beta + \left\lfloor \frac{n+\alpha-\beta}{2} \right\rfloor\right\},$$

where $\lfloor x \rfloor$ denotes that the largest integer smaller or equal to x. Let $L_0 = \left\{\alpha, \alpha+1, \ldots, \lfloor \frac{n+\beta-\alpha}{2} \rfloor + \alpha\right\}$, $L_1 = \left\{\beta, \beta+1, \ldots, \lfloor \frac{n+\alpha-\beta}{2} \rfloor + \beta\right\}$. There are three possibilities: (1) $\alpha - \lfloor\alpha\rfloor \neq \beta - \lfloor\beta\rfloor$; (2) $\alpha - \lfloor\alpha\rfloor = \beta - \lfloor\beta\rfloor$ and $\alpha = \beta$; (3) $\alpha - \lfloor\alpha\rfloor = \beta - \lfloor\beta\rfloor$ but $\alpha \neq \beta$. We will provide the derivation only in case (3); the other cases are similar, and lead to the exact same expressions.

We assume that $\alpha - \lfloor\alpha\rfloor = \beta - \lfloor\beta\rfloor$ but $\alpha \neq \beta$. Without loss of generality, we assume that $\alpha > \beta$, and let $\alpha = \beta + \delta$, where δ is a positive integer. Notice that it is easy to show in this case that $\lfloor \frac{n+\beta-\alpha}{2} \rfloor + \alpha = \lfloor \frac{n+\alpha-\beta}{2} \rfloor + \beta$ by considering the evenness and oddness of n and δ. Therefore, we have $L_0 \subset L_1$. In the following, we discuss two cases when $n+\beta-\alpha$ is even and when $n+\beta-\alpha$ is odd.

(1) When $n+\beta-\alpha$ is even, the event $[M = m]$ is equal to the union of the disjoint events $[n_0 = m-\alpha]$, and $[n_1 = m-\beta] = [n_0 = n-m+\beta]$, for $m \in L_0 \setminus \left\{\alpha + \frac{n+\beta-\alpha}{2}\right\}$, whereas $\left[M = \alpha + \frac{n+\beta-\alpha}{2}\right] = \left[n_0 = m-\alpha = \frac{n+\beta-\alpha}{2}\right]$, and $[M = m] = [n_0 = n-m+\beta]$, for $m \in L_1 \setminus L_0$.

Now, we are going to use the fact that, for a random variable X and disjoint events A and B, one has[1]

$$E\left[X \mid A \cup B\right] = \frac{P(A)}{P(A) + P(B)} E\left[X \mid A\right] + \frac{P(B)}{P(A)+P(B)} E[X \mid B]\,. \tag{25}$$

We can then write $E\left[\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{\hat{\varepsilon}_{0,\mathrm{OBP}}}\right]$ as:

$$\begin{aligned}
E\left[\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{\hat{\varepsilon}_{0,\mathrm{OBP}}}\right] &= \sum_{m\in L_0\setminus\left\{\frac{n+\Delta_c}{2}\right\}} E\left[\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{m/(n+\Delta_c)} \;\middle|\; n_0 = m-\alpha\right]P(n_0 = m-\alpha) + \\
&\quad \sum_{m\in L_0\setminus\left\{\frac{n+\Delta_c}{2}\right\}} E\left[\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{m/(n+\Delta_c)} \;\middle|\; n_0 = n-m+\beta\right]P(n_0 = n-m+\beta) + \\
&\quad \sum_{m\in L_1\setminus L_0} E\left[\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{m/(n+\Delta_c)} \;\middle|\; n_0 = n-m+\beta\right]P(n_0 = n-m+\beta) + \\
&\quad E\left[\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{m/(n+\Delta_c)} \;\middle|\; n_0 = \frac{n-\alpha+\beta}{2}\right]P\left(n_0 = \frac{n-\alpha+\beta}{2}\right) \\
&= \sum_{m\in L_0} E\left[\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{m/(n+\Delta_c)} \;\middle|\; n_0 = m-\alpha\right]P(n_0 = m-\alpha) + \\
&\quad \sum_{m\in L_1} E\left[\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{m/(n+\Delta_c)} \;\middle|\; n_0 = n-m+\beta\right]P(n_0 = n-m+\beta)I_{m\neq\frac{n+\alpha+\beta}{2}} \\
&= \sum_{r=0}^{\lfloor\frac{n+\beta-\alpha}{2}\rfloor} E\left[\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{(n_0+\alpha)/(n+\Delta_c)} \;\middle|\; n_0 = r\right]P(n_0 = r) + \\
&\quad \sum_{r=0}^{\lfloor\frac{n+\alpha-\beta}{2}\rfloor} E\left[\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{(n_1+\beta)/(n+\Delta_c)} \;\middle|\; n_0 = n-r\right]P(n_0 = n-r)I_{r\neq\frac{n+\alpha-\beta}{2}}
\end{aligned} \tag{26}$$

where $P(n_0 = r) = \binom{n}{r}c^r(1-c)^{n-r}$ and

$$\begin{aligned}
&E\left[\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{(n_0+\alpha)/(n+\Delta_c)} \;\middle|\; n_0 = r\right] \\
&= \sum_{i=1}^{2^d}\left\{\sum_{\substack{\frac{(r+\alpha)(k+\alpha_i)}{r+\Delta_p} < \frac{(n-r+\beta)(l+\beta_i)}{n-r+\Delta_q} \\ k\le r,\ k+l\le n}} \frac{(r+\alpha)(k+\alpha_i)}{r+\Delta_p}P(U_i = k, V_i = l \mid n_0 = r)\right. \\
&\quad \left. + \sum_{\substack{\frac{(r+\alpha)(k+\alpha_i)}{r+\Delta_p} \ge \frac{(n-r+\beta)(l+\beta_i)}{n-r+\Delta_q} \\ k\le r,\ k+l\le n}} \frac{(n-r+\beta)(l+\beta_i)}{n-r+\Delta_q}P(U_i = k, V_i = l \mid n_0 = r)\right\},
\end{aligned} \tag{27}$$

with $P(U_i = k, V_i = l \mid n_0 = r) = \binom{r}{k}p_i^k(1-p_i)^{r-k}\binom{n-r}{l}q_i^l(1-q_i)^{n-r-l}$. The expression for $E\left[\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{(n_1+\beta)/(n+\Delta_c)} \mid n_0 = n-r\right]$ is obtained from (27), with $r+\alpha$ replaced by $r+\beta$.

(2) When $n+\beta-\alpha$ is odd, the event $[M = m]$ is equal to the union of the disjoint events $[n_0 = m-\alpha]$, and $[n_1 = m-\beta] = [n_0 = n-m+\beta]$, for $m \in L_0$, whereas $[M = m] = [n_0 = n-m+\beta]$, for $m \in L_1\setminus L_0$. By applying the same reasoning, we have the same expression as in (26). Note that $I_{n_1\neq\frac{n+\alpha-\beta}{2}}$ is always equal to 1 in this case since $\frac{n+\alpha-\beta}{2}$ is not an integer.

For the second moment, we have

$$E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBP}}}{\hat{\varepsilon}^2_{0,\mathrm{OBP}}}\right] = E\left[E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBP}}}{\hat{\varepsilon}^2_{0,\mathrm{OBP}}} \;\middle|\; \hat{\varepsilon}_{0,\mathrm{OBP}}\right]\right] = \sum_{m\in L} E\left[\left(\frac{\hat{\varepsilon}_{\mathrm{OBP}}}{m/(n+\Delta_c)}\right)^2 \;\middle|\; M = m\right]P(M = m), \tag{28}$$

where $M = (n + \Delta_c)\,\hat{\varepsilon}_{0,\mathrm{OBP}}$, as before. By using the same reasoning applied previously in the case of the first moment, we have

$$\begin{aligned}
E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBP}}}{\hat{\varepsilon}^2_{0,\mathrm{OBP}}}\right] &= \sum_{m\in L_0} E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBP}}}{m^2/(n+\Delta_c)^2}\;\middle|\; n_0 = m-\alpha\right] \\
&\quad P(n_0 = m-\alpha) + \\
&\quad \sum_{m\in L_1} E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBP}}}{m^2/(n+\Delta_c)^2}\;\middle|\; n_0 = n-m+\beta\right] \\
&\quad P(n_0 = n-m+\beta) I_{m\neq\frac{n+\alpha+\beta}{2}}, \\
&= \sum_{t=0}^{\lfloor\frac{n+\beta-\alpha}{2}\rfloor} E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBP}}}{(n_0+\alpha)^2/(n+\Delta_c)^2}\;\middle|\; n_0 = t\right] \\
&\quad P(n_0 = t) + \\
&\quad \sum_{t=0}^{\lfloor\frac{n+\alpha-\beta}{2}\rfloor} E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBP}}}{(n_1+\beta)^2/(n+\Delta_c)^2}\;\middle|\; n_0 = n-t\right] \\
&\quad P(n_0 = n-t) I_{t\neq\frac{n+\alpha-\beta}{2}}
\end{aligned} \tag{29}$$

where, letting $k'_i = \frac{(r+\alpha)(k+\alpha_i)}{r+\alpha_i}$, $l'_i = \frac{(n-r+\alpha)(l+\beta_i)}{n-r+\beta_i}$, $r'_j = \frac{(r+\alpha)(r+\alpha_i)}{r+\alpha_j}$, $s'_j = \frac{(n-r+\alpha)(s+\beta_i)}{n-r+\beta_j}$, for $i,j = 1,\ldots,b$, we have

$$\begin{aligned}
&E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBP}}}{(n_0+\alpha)^2/(n+\Delta_c)^2}\;\middle|\; n_0 = t\right] = \frac{1}{(t+\alpha)^2}\times \\
&\sum_{i=1}^{2^d}\left\{\sum_{l'_i>k'_i} k'^2_j P(U_i = k'_i, V_i = l'_i \mid n_0 = t)\right. \\
&\qquad \left. + \sum_{k\geq l} l'^2_i P(U_i = k'_i, V_i = l'_i \mid n_0 = t)\right\} + \\
&\frac{1}{(t+\alpha)^2}\sum_{\substack{i,j=1\\ i\neq j}}^{2^d}\left\{\sum_{l'_i>k'_i}\sum_{s'_j>r'_j} k'_i r'_j P(U_i = k'_i, V_i = l'_i, U_j = r'_j,\right. \\
&\qquad V_j = s'_j \mid n_0 = t) + \\
&\qquad \sum_{l'_i>k'_i}\sum_{r'_j\geq s'_j} k'_i s'_j P(U_i = k'_i, V_i = l'_i, U_j = r'_j, \\
&\qquad V_j = s'_j \mid n_0 = t) + \\
&\qquad \sum_{k'_i\geq l'_i}\sum_{s'_j>r'_j} l'_i r'_j P(U_i = k'_i, V_i = l'_i, U_j = r'_j, \\
&\qquad V_j = s'_j \mid n_0 = t) \; + \\
&\qquad \sum_{k'_i\geq l'_i}\sum_{r'_j\geq s'_j} l'_i s'_j P(U_i = k'_i, V_i = l'_i, U_j = r'_j, \\
&\qquad \left. V_j = s'_j \mid n_0 = t)\right\},
\end{aligned} \tag{30}$$

with $P(U_i = k, V_i = l, U_j = r, V_j = s \mid n_0 = t) = \binom{n_0}{k,r} p_i^k p_j^r (1-p_i-p_j)^{n_0-k-r} \binom{n-n_0}{l,s} q_i^l q_j^s (1-q_i-q_j)^{n-n_0-l-s}$. The expression for $E\left[\frac{\hat{\varepsilon}^2_{\mathrm{OBP}}}{(n_1+\beta)^2/(n+\Delta_c)^2}\;\middle|\; n_0 = n-t\right]$ is obtained from (30) with $t+\alpha$ replaced by $t+\beta$.

5 Numerical experiments

5.1 Global accuracy

In this section, we employ Monte Carlo sampling (with $M = 10,000$ simulated data sets for each sample size) to compute global accuracy metrics of the two Bayesian CoD estimators. Following [17], we let $\alpha = 2^d+1 = \beta = 2^d+1$, which produces a prior for c peaked around the value $c = 0.5$, and $\alpha_i = \beta_i = 1$, for all $i = 1,\ldots,2^d$, i.e. flat (uniform) prior distributions for $(\mathbf{p},\mathbf{q})$. In each iteration, the values of c and $(\mathbf{p},\mathbf{q})$ are drawn from the respective priors, and then sample data is generated according to these probabilities. Given the sample data, we compute the exact Bayesian MMSE and OBP CoD estimates as expressed in Section 3, and compare them to the standard resubstitution CoD estimator, which is based on plugging in sample frequencies in the expression for the optimal CoD, and corresponds to the the original choice of CoD estimator in [7]. This estimator is also called the nonparametric maximum-likelihood CoD estimator in [15]. For further comparison, we also compute CoD estimators based on leave-one-out, 0.632 bootstrap and 10-repeated twofold cross-validation error estimators—for details on all these CoD estimators, please see [14, 15]. Sample means and sample variances are employed to approximate the global accuracy metrics of each CoD estimator.

Figure 1 displays the global bias, variance, and RMS as a function of varying sample size, for different numbers of binary predictive variables, $d = 1$ through $d = 3$. Several observations are evident. First, as expected, the Bayesian MMSE CoD estimator is unbiased and has the least RMS among all the estimators, and the gap in performance widens as dimensionality increases. Secondly, the OBP CoD estimator has the second-best performance, which indicates the benefits of using the Bayesian estimation approach. The accuracy of the OBP CoD estimator is quite close to that of the MMSE estimator for $d = 1$, but the gap widens as d increases. Thirdly, it is also observed that the OBP Bayesian CoD estimator is pessimistically biased. Incidentally, the 0.632 bootstrap CoD estimator displays the best accuracy among the four nonparametric ones according to global RMS, but it is matched by the resubstitution CoD estimator as sample size increases.

5.2 Fixed-parameter accuracy

In this section, we study the average accuracy of the two proposed Bayesian CoD estimators for a fixed parameter, that is, we evaluate the Bayesian estimators from a purely frequentist perspective.

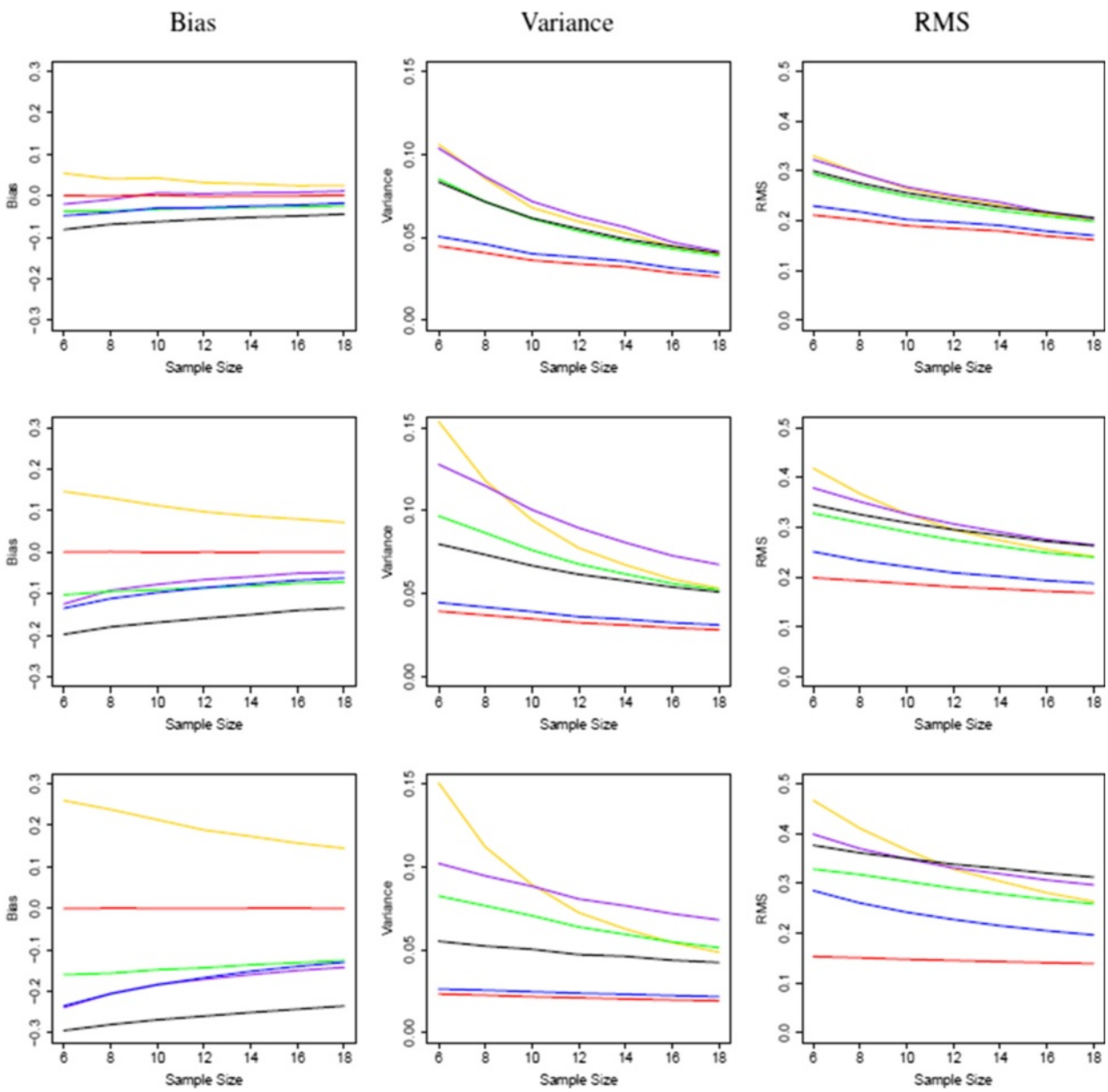

Fig. 1 Global bias, variance, and RMS for several CoD estimators versus sample size n, for different numbers of predicting binary variables d. *Top row*: $d = 1$; *Middle row*: $d = 2$; *Bottom row*: $d = 3$. Plot key: MMSE (*red*), OBP (*blue*), resubstitution (*gold*), leave-one-out (*purple*), 0.632 bootstrap (*green*), 10-repeated twofold cross-validation (*black*)

As in the previous subsection, we consider $d = 1$ through $d = 3$ binary predictive variables. We consider fixed values of the parameters, c^*, $\mathbf{p}^*$, and $\mathbf{q}^*$. In order to examine the effect of prior belief on performance, we consider four scenarios regarding prior density around the true parameters: a flat ("non-informative") prior and three nonflat "matched," "poorly matched," and "mismatched" priors. This is done by assuming different base measures and concentration parameters for the priors (c.f. Section 3). As an illustration of the approach, consider the case $d = 1$. In our simulation, $c^* = 0.5$, $\mathbf{p}^* = (0.6, 0.4)$, and $\mathbf{q}^* = (0.4, 0.6)$, and the base measures for the nonflat priors are $c_0 = 0.5$, $\mathbf{p}_0 = \mathbf{p}^*$, $\mathbf{q}_0 = \mathbf{q}^*$ (matched prior), $c_0 = 0.5$, $\mathbf{p}_0 = (0.5, 0.5)$, $\mathbf{q}_0 = (0.5, 0.5)$ (poorly matched prior), and $c_0 = 0.5$, $\mathbf{p}_0 = (0.4, 0.6)$, $\mathbf{q}_0 = (0.6, 0.4)$ (mismatched prior). In addition, we consider different values of the concentration parameters to reflect different degrees of peaking of the prior distributions. Figure 2 plots the nonflat prior densities for p_1 (which is Beta-distributed), for different values of the concentration

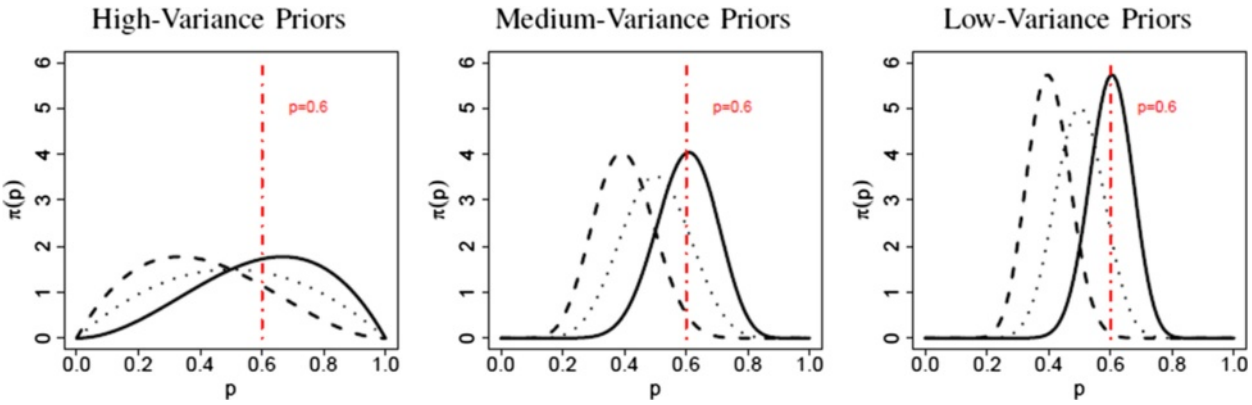

Fig. 2 Beta prior densities for p_1 and $d = 1$, for different values of the concentration parameter: $\Delta_p = 5$ (high-variance priors), $\Delta_p = 25$ (medium-variance priors), and $\Delta_p = 50$ (low-variance priors). Legend: *dashed line* (mismatched prior), *dotted line* (poorly matched prior), and *solid line* (matched prior). The base measures are given in Table 1, and the true value of the parameter is indicated by a *vertical line*

parameter: $\Delta_p = 5$ (high-variance), $\Delta_p = 25$ (medium variance), and $\Delta_p = 50$ (low variance). Notice that each density is centered around the expected value. Note that, if the variance is high, even the matched prior becomes very diffuse around its expected value (which is the true value, in this case).

Table 1 gives the values of the parameters used in the experiments. In all cases, the true value and base measure for c are the same, $c^* = c_0 = 0.5$. In addition, in each case, the true value $\mathbf{q}^*$ and base measure $\mathbf{q}_0$ are obtained from $\mathbf{p}^*$ and $\mathbf{p}_0$, respectively, by flipping the corresponding vector left to right; for example, when $\mathbf{p}_0 = (0.2, 0.1, 0.3, 0.4)$ then $\mathbf{q}_0 = (0.4, 0.3, 0.1, 0.2)$. Therefore, only the values for $\mathbf{p}^*$ and $\mathbf{p}_0$ are shown in Table 1.

Figures 3, 4, and 5 show the results for $d = 1$ through $d = 3$ predictors, respectively. Each figure displays the bias, variance, and RMS as a function of the sample size for the Bayesian MMSE and OBP CoD estimators and the nonparametric CoD estimators. The Bayesian estimators assume a flat non-informative prior and three nonflat matched, poorly matched, and mismatched priors, specified by Table 1. For the non-flat priors only, three different variance groups are considered, corresponding to three different settings for the concentration parameters: high variance, medium variance, and low variance priors. Results for the OBP CoD estimator are computed exactly using the results of Section 4. For all other CoD estimators, bias, variance, and RMS are approximated by averaging results over 5000 Monte Carlo samples drawn from the fixed distribution. The curves for the nonparametric CoD estimators and the flat prior Bayesian CoD estimators are repeated across the columns, for comparison with the results for the nonflat prior Bayesian CoD estimators.

We can observe that, as expected, both Bayesian CoD estimators perform better when the prior is matched to the true value of the parameters than when the match is poor or nonexistent. In addition, for the matched prior, accuracy improves substantially as one moves from a diffuse (high-variance) to a peaked (low-variance) prior. This effect is especially visible in the case of the OBP CoD estimator. For example, with $d = 1$ the RMS is reduced by nearly 80 % between the high-variance and low-variance matched priors. In fact, the accuracy of the OBP CoD estimator beats that of the MMSE CoD estimator for peaked priors, while the opposite is true under diffuse priors. Both Bayesian CoD estimators outperform the nonparametric ones in cases $d = 1$ and $d = 3$, whereas, in the

Table 1 True distributions and nonflat prior base measures for fixed-parameter experiments. In all cases, $c^* = c_0 = 0.5$, and $\mathbf{q}^*$ and $\mathbf{q}_0$ are obtained from $\mathbf{p}^*$ and $\mathbf{p}_0$, respectively, by flipping left to right (see text.)

	True distribution	Base measure 1	Base measure 2	Base measure 3
$d = 1$	$\mathbf{p}^* = (0.6, 0.4)$	$\mathbf{p}_0^1 = (0.6, 0.4)$	$\mathbf{p}_0^2 = (0.5, 0.5)$	$\mathbf{p}_0^3 = (0.4, 0.6)$
$d = 2$	$\mathbf{p}^* = (0.2, 0.3, 0.1, 0.4)$	$\mathbf{p}_0^1 = (0.2, 0.3, 0.1, 0.4)$	$\mathbf{p}_0^2 = (0.3, 0.2, 0.2, 0.3)$	$\mathbf{p}_0^3 = (0.4, 0.1, 0.3, 0.2)$
$d = 3$	$\mathbf{p}^* = (0.1, 0.15, 0.05, 0.2, 0.15, 0.1, 0.1, 0.15)$	$\mathbf{p}_0^1 = (0.1, 0.15, 0.05, 0.2, 0.15, 0.1, 0.1, 0.15)$	$\mathbf{p}_0^2 = (0.15, 0.1, 0.1, 0.15, 0.1, 0.05, 0.15, 0.2)$	$\mathbf{p}_0^3 = (0.2, 0.05, 0.15, 0.1, 0.05, 0.2, 0.2, 0.05)$
		Matched prior	Poorly matched prior	Mismatched prior

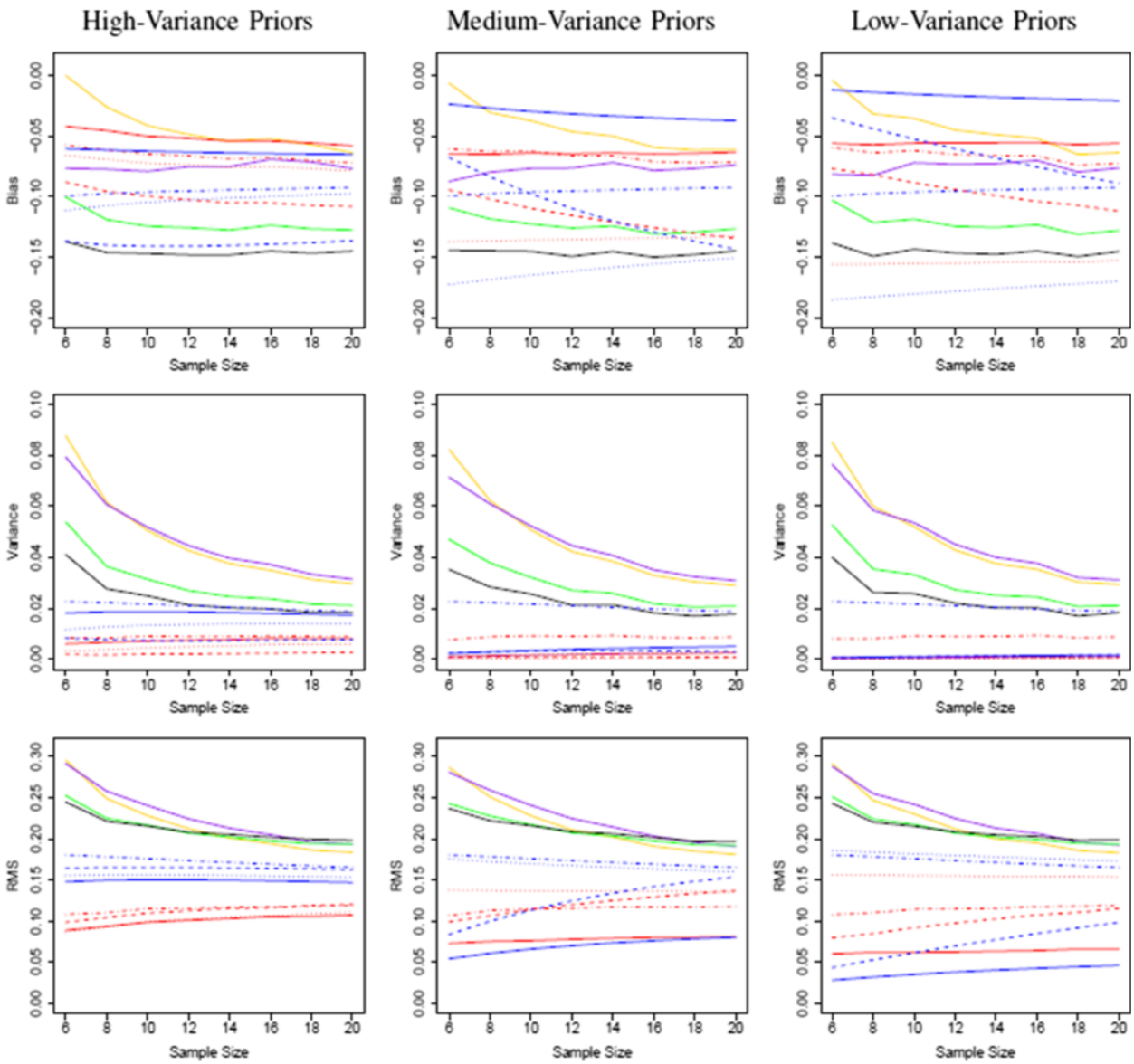

Fig. 3 Bias, variance, and RMS for several CoD estimators versus sample size for $d = 1$, under different values of the concentration parameter: $\Delta_c/2 = \Delta_p = \Delta_q = 5$ (high-variance priors), $\Delta_c/2 = \Delta_p = \Delta_q = 25$ (medium-variance priors), and $\Delta_c/2 = \Delta_p = \Delta_q = 50$ (low-variance priors). Plot key for base measures: *dash-dot line* (uniform prior), *dashed line* (mismatched prior), *dotted line* (poorly matched prior), and *solid line* (matched prior); see Table 1. Plot key for CoD estimators: MMSE (*red*), OBP (*blue*), resubstitution (*gold*), leave-one-out (*purple*), 0.632 bootstrap (*green*), 10-repeated twofold cross-validation (*black*). Results for the OBP CoD estimator are exact while others are approximated by Monte Carlo sampling method. The curves for the nonparametric CoD estimators and the flat prior Bayesian CoD estimators are repeated across the columns, for comparison with the results for the nonflat prior Bayesian CoD estimators

$d = 2$ case, the Bayesian estimators based on mismatched or poorly matched priors can perform worse than the nonparametric estimators, for larger sample size. It is also observed that, as the variance of priors decreases (i.e., for a larger Δ value), the performance of both Bayesian estimators improves over the nonparametric ones. Moreover, it is interesting that the Bayesian MMSE CoD estimator performs better than the OBP CoD estimator, for a high-variance prior with matched prior, while the OBP CoD estimator beats the Bayesian MMSE CoD estimator for medium and high-variance matched priors. This indicates that the OBP CoD estimator is preferable due to its straightforward representation and superior performance with low-variance priors. Notice that the Bayesian MMSE CoD estimator has the least RMS only when averaged over all distributions and all possible samples, but

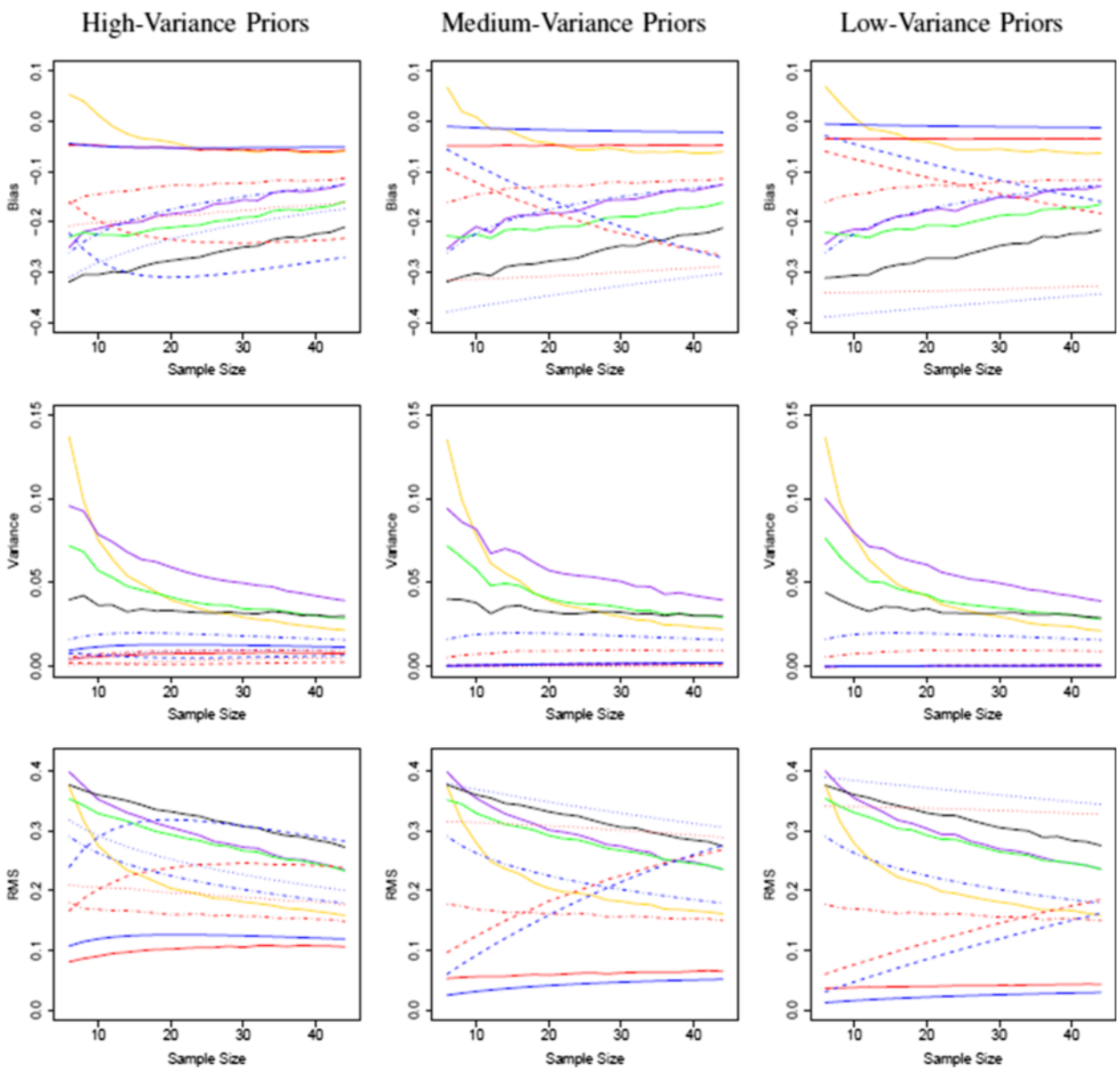

Fig. 4 Bias, variance, and RMS for several CoD estimators versus sample size for $d = 2$, under different values of the concentration parameter: $\Delta_c/2 = \Delta_p = \Delta_q = 10$ (high-variance priors), $\Delta_c/2 = \Delta_p = \Delta_q = 50$ (medium-variance priors), and $\Delta_c/2 = \Delta_p = \Delta_q = 100$ (low-variance priors). Plot key for base measures: *dash-dot line* (uniform prior), *dashed line* (mismatched prior), *dotted line* (poorly matched prior), and *solid line* (matched prior); see Table 1. Plot key for CoD estimators: MMSE (*red*), OBP (*blue*), resubstitution (*gold*), leave-one-out (*purple*), 0.632 bootstrap (*green*), 10-repeated twofold cross-validation (*black*). Results for the OBP CoD estimator are exact while others are approximated by Monte Carlo sampling method. The curves for the nonparametric CoD estimators and the flat prior Bayesian CoD estimators are repeated across the columns, for comparison with the results for the nonflat prior Bayesian CoD estimators

its optimality does not apply to the settings with a fixed distribution. In addition, we observe that the Bayesian MMSE CoD estimator is less variant than the OBP CoD estimator. It can be seen that the Bayesian CoD estimators based on informative priors are less variant than those based on non-informative uniform priors. In the $d = 1$ and $d = 3$ cases, the OBP CoD estimator with uniform priors becomes more variant than even the cross-validation estimator, for larger sample size. In addition, the OBP CoD estimator is less biased in magnitude than the MMSE estimator for low-variance matched priors. However, as the variance of priors increases, the Bayesian MMSE CoD estimator turns out to have less bias than the OBP estimator.

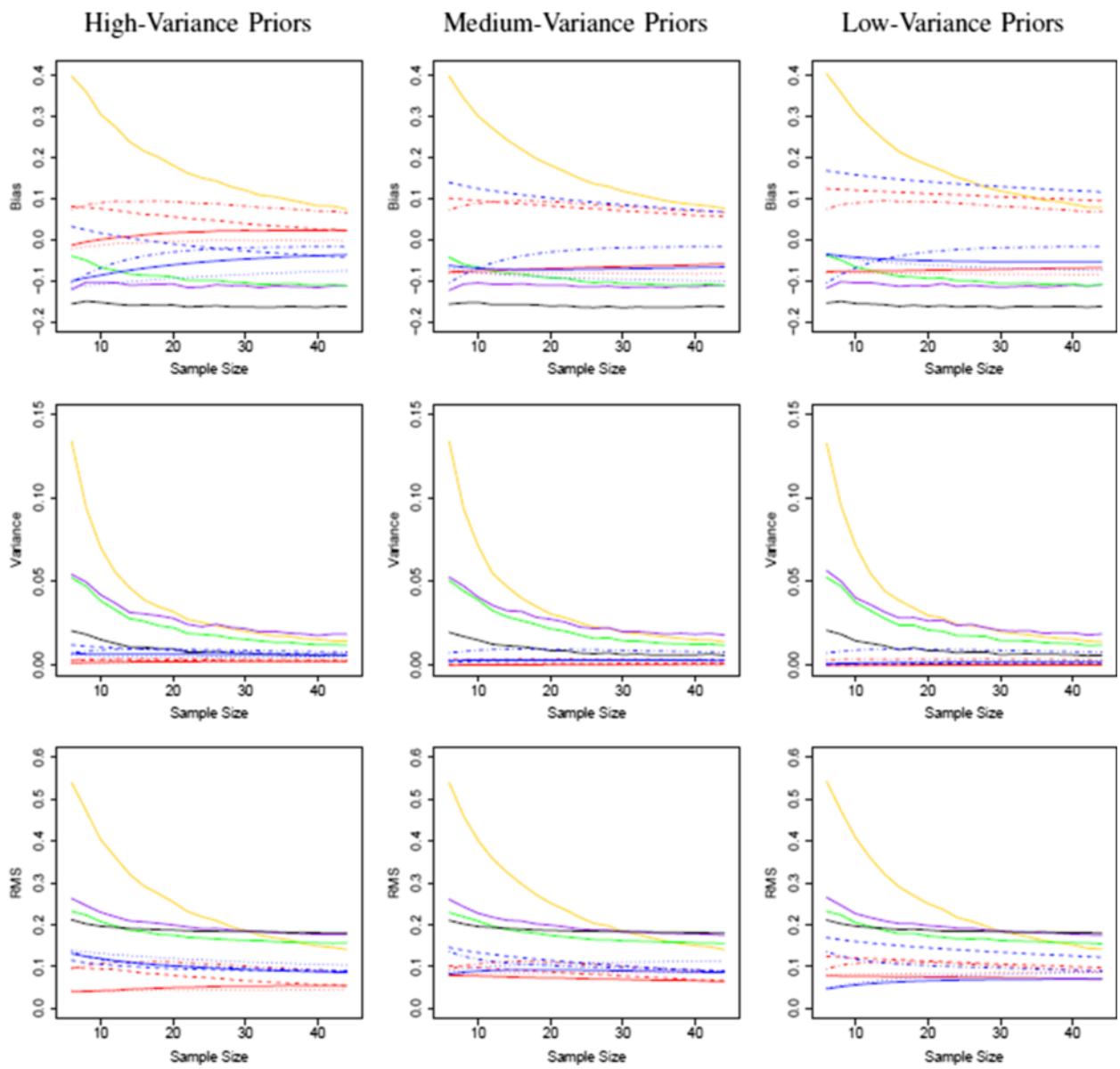

Fig. 5 Bias, variance, and RMS for several CoD estimators versus sample size for $d = 3$, under different values of the concentration parameter: $\Delta_c/2 = \Delta_p = \Delta_q = 20$ (high-variance priors), $\Delta_c/2 = \Delta_p = \Delta_q = 100$ (medium-variance priors), and $\Delta_c/2 = \Delta_p = \Delta_q = 200$ (low-variance priors). Plot key for base measures: *dash-dot line* (uniform prior), *dashed line* (mismatched prior), *dotted line* (poorly matched prior), and *solid line* (matched prior); see Table 1. Plot key for CoD estimators: MMSE (*red*), OBP (*blue*), resubstitution (*gold*), leave-one-out (*purple*), 0.632 bootstrap (*green*), 10-repeated twofold cross-validation (*black*). Results for the OBP CoD estimator are exact while others are approximated by Monte Carlo sampling method. The curves for the nonparametric CoD estimators and the flat prior Bayesian CoD estimators are repeated across the columns, for comparison with the results for the nonflat prior Bayesian CoD estimators

6 Gene regulatory network inference: a melanoma example

We discuss in this section the application of the Bayesian CoD estimation approach discussed previously to the inference of gene regulatory networks. We apply the proposed inference procedure on data collected in a study of metastatic melanoma [24], containing 31 binarized sample expression profiles, which have been binarized, with 0 indicate no significant expression whereas 1 represents significant expression (either over- or under-expression). It was found in [24] that the WNT5A gene is a major driver of processes that lead to metastatic melanoma. We derive the logic relationships and wiring of a 7-gene WNT5A network consisting of genes selected using data analysis and prior biological knowledge: WNT5A, pirin, S100P, RET1, MART1, HADHB, and

STC2; for more details about the selection of these genes, see [25, 26].

We assume a model where the target binary gene expression $Y \in \{0,1\}$ is regulated by a binary predictor gene expression vector $\mathbf{X} = (X_1, \ldots, X_d) \in \{0,1\}^d$ through the relationship

$$Y = f(\mathbf{X}) \oplus N, \tag{31}$$

where $f : \{0,1\}^d \to \{0,1\}$ is a Boolean function, the symbol "$\oplus$" indicates modulo-2 addition, and $N \in \{0,1\}$ is a noise Bernoulli random variable, independent from $\mathbf{X}$, such that $P(N = 0) = p$. The modulo-2 addition behaves as a XOR operation, which flips the state of the target Y when $N = 1$, and leaves it unaltered when $N = 0$. Hence, p quantifies the predictive power of the model: if $p = 1$, the system is noiseless and prediction is deterministic, while if $p < 1$, there is a degree of indeterminacy in the state of the target given the state of the predictors. This model is studied in detail in [15], where an inference procedure, based on a maximum-likelihood CoD estimator, is proposed to select the unknown Boolean function f, assuming that f is a member of a candidate model set F containing Boolean functions that depend on the same number k of essential variables. Each f in F is specified by (1) a Boolean function $g : \{0,1\}^k \to \{0,1\}$ and (2) the indices for the predicting variable set $\{i_1, \ldots, i_k\} \subset \{1, \ldots, d\}$, or *wiring*, such that $f(\mathbf{X}) = g\left(X_{i_1}, \ldots, X_{i_k}\right)$. If the candidate boolean functions g belong to a model set G, then the total number of possible models is $|G| \times \binom{d}{l}$.

Here, we modify the network inference in [15] to allow the use of the Bayesian CoD estimators described previously. For a given target Y and predictor set $\mathbf{X}$, we assume Dirichlet prior distributions as in (5). Instead of adopting a non-informative choice of hyperparameters, we employ an "empirical Bayes" approach, where the hyperparameters are estimated in part from the sample data, as described next.

First, it follows from the model in (31) that the parameters $p_i = P(\mathbf{X} = \mathbf{x}^i \mid Y = 0)$, $q_i = P(\mathbf{X} = \mathbf{x}^i \mid Y = 1)$, and $c = P(Y = 0)$ are given by:

$$\begin{aligned} p_i &\propto \left(p(1 - f(\mathbf{x}^i)) + (1-p) f(\mathbf{x}^i)\right) P(\mathbf{X} = \mathbf{x}^i), \quad i = 1, \ldots, 2^d, \\ q_i &\propto \left(p f(\mathbf{x}^i) + (1-p)(1 - f(\mathbf{x}^i))\right) P(\mathbf{X} = \mathbf{x}^i), \quad i = 1, \ldots, 2^d, \\ c &= \textstyle\sum_{i=1}^{2^d} \left(p(1 - f(\mathbf{x}^i)) + (1-p) f(\mathbf{x}^i)\right) P(\mathbf{X} = \mathbf{x}^i). \end{aligned} \tag{32}$$

The unknown quantities here are the predictive power p and the distribution $P(\mathbf{X})$ of the predictors. Given the sample data $\mathbf{S}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, and a fixed Boolean function g and wiring $\{i_1, \ldots, i_k\}$, p can be very effectively estimated by means of the sample frequency [15]:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} I_{f(\mathbf{X}_i) = Y_i}. \tag{33}$$

The distribution $P(\mathbf{X})$ could in principle be also estimated from the data using sample frequencies; however, such an estimator can become very poor under small sample sizes and large dimensionality d. Therefore, we simply assume a flat distribution $P(\mathbf{X} = \mathbf{x}^i) = 1/2^d$, for $i = 1, \ldots, 2^d$. Substituting this and (33) into (32) gives the values of the hyperparameters used in our experiment:

$$\begin{aligned} \hat{p}_i &\propto \left(\hat{p}(1 - f(\mathbf{x}^i)) + (1 - \hat{p}) f(\mathbf{x}^i)\right), \quad i = 1, \ldots, 2^d \\ \hat{q}_i &\propto \left(\hat{p} f(\mathbf{x}^i) + (1 - \hat{p})(1 - f(\mathbf{x}^i))\right), \quad i = 1, \ldots, 2^d \\ \hat{c} &= (1/2^d) \textstyle\sum_{i=1}^{2^d} \left(\hat{p}(1 - f(\mathbf{x}^i)) + (1 - \hat{p}) f(\mathbf{x}^i)\right). \end{aligned} \tag{34}$$

Recall from Section 3 that the shape of the Dirichlet prior distribution is determined by the hyperparameters through a location parameter, called the base measure, and a concentration parameter. Our strategy is to set up the estimates in (34) as the base measure, so that the Dirichlet priors are concentrated around them, to a degree specified by the concentration parameter. Formally, the hyperparameters are set to: $\{\alpha_1, \ldots, \alpha_{2^d}\} = \{\lceil \hat{p}_1 \Delta \rceil, \ldots, \lceil \hat{p}_{2^d} \Delta \rceil\}$, $\{\beta_1, \ldots, \beta_{2^d}\} = \{\lceil \hat{q}_1 \Delta \rceil, \ldots, \lceil \hat{q}_{2^d} \Delta \rceil\}$, $\alpha = \lceil \hat{c} \Delta \rceil$ and $\beta = \lceil (1 - \hat{c}) \Delta \rceil$, where $\lceil x \rceil$ gives the smallest integer larger or equal to x. The value of Δ is tuned by the experimenter, either manually or using a data-driven procedure.

We are now ready to state the procedure to select a function f in F, consisting of a k-predictor Boolean function g and its wiring.

6.1 Bayesian model selection procedure

1. For each of the Boolean functions $g \in G$, compute the prior hyperparameters as described earlier. Obtain the MMSE Bayesian CoD / OBP CoD estimate under each of the $\binom{d}{k}$ possible wirings. Pick the wiring for g that produces the largest CoD estimate. Ties, if any, are broken randomly.
2. Among the $|G|$ pairs of Boolean function g and wiring obtained in the previous step, select the one that produces the largest predictive power estimate $\hat{p}$. Ties, if any, are broken randomly.

In our experiment with the 7-gene WNT5A network, we consider in turn each gene as a target and the remaining six genes as predictors (so that a gene cannot be a predictor of itself). Hence, $d = 6$. In addition, we assume that that each gene is predicted by three genes out of the

six predictors. Therefore, $k = 3$ and there are $\binom{6}{3} = 20$ possible wirings for each target gene. The set G contains all 218 Boolean functions of exactly three essential variables (this is less than the full set of $2^{2^3} = 256$ 3-input Boolean functions since those that are reducible to 0-, 1-, and 2-input logics are not considered). We set $\Delta = 1.0$ and apply the proposed Bayesian model selection procedure to infer a gene regulatory network for the MMSE and OBP CoD estimators. We also obtain the gene regulatory network produced by employing the standard model selection procedure, which picks the predictor set (among all $\binom{6}{3} = 20$ choices, in this case) with the largest estimated resubstitution CoD [25].

The results are presented in Fig. 6. The diagrams represent the predicted logic functions as binary strings (in the usual logic table order; e.g., AND = 00000001) and the predicted wirings as oriented edges, and, in addition, the estimated CoD in each case is displayed. We can see that the predicted logic functions and wirings for the three networks are similar, especially in the cases of the OBP and resubstitution CoD networks. If one considers only the three top predicted relationships according to CoD magnitude, one obtains the diagrams depicted in Fig. 7, which show that the same network is inferred by the OBP and resubstitution CoDs, which differ from the network obtained with the MMSE CoD by only a single arrow shift in the wirings (the inferred logics in all three cases are also very similar, differing by only a few bit shifts). The important difference between the Bayesian and standard approaches that can be observed from this experiment is in the estimated CoD magnitudes: those estimated with the standard resubstitution CoD tend to be much larger than the ones estimated with the Bayesian CoDs. This reflects the optimistic bias that tends to be displayed by resubstitution [27], a problem that is avoided by the Bayesian CoD estimators.

7 Conclusions

We introduced a Bayesian framework for the estimation of the CoD in a discrete prediction setting and analyzed the accuracy of the proposed Bayesian MMSE and OBP CoD estimators based on fixed and random parameters, using analytical and simulation methods. We also compared the accuracy of the two Bayesian CoD estimators against those of several classical CoD estimators, based on resubstitution, leave-one-out, bootstrap, and cross-validation prediction error estimation. Our results indicated that the Bayesian MMSE CoD estimator has the best performance with zero bias and least RMS, when averaged over all distributions and sample data, whereas, for fixed distributions, we conclude that priors with higher densities around the fixed distributions present better accuracy with less RMS. It is also interesting to see that the OBP CoD estimator, one with very simple calculation, can beat the Bayesian MMSE CoD estimator when using low-variance priors with higher densities around the parameters of the fixed distributions. Furthermore, we proposed an approach for inference of gene regulatory networks based on the proposed Bayesian CoD estimators, and applied it to the inference of a 7-gene regulatory network using melanoma data. We observed that the inferred boolean functions and wirings were similar for both CoD Bayesian estimators. Interestingly, the network inferred with the OBP CoD estimator was very close to the network obtained with the standard inference method based on the resubstitution CoD estimator; however, the magnitude of the CoDs were larger in the latter case, which is consistent with the fact that resubstitution tends to be optimistic. We hope that this paper will provide a theoretical foundation for further work on Bayesian estimation methodologies for the inference of gene regulatory networks. The issue of obtaining informative priors based on established biological knowledge about regulatory relationships, which was not addressed in detail here, is one that deserves careful consideration in future work on this topic.

Endnote

[1] A proof of this fact is given in the Appendix of [14].

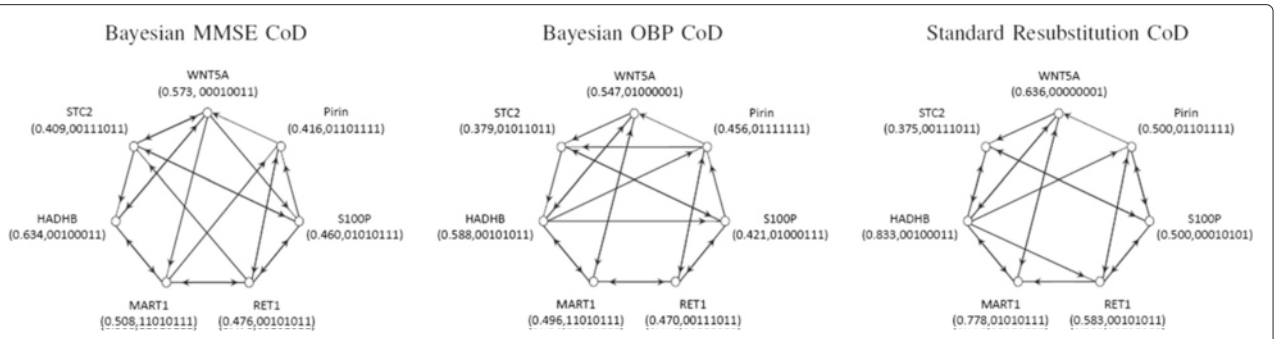

Fig. 6 Gene regulatory networks inferred using the Bayesian model selection procedure for the Bayesian MMSE CoD and OBP CoD and the standard model selection procedure using the resubstitution CoD

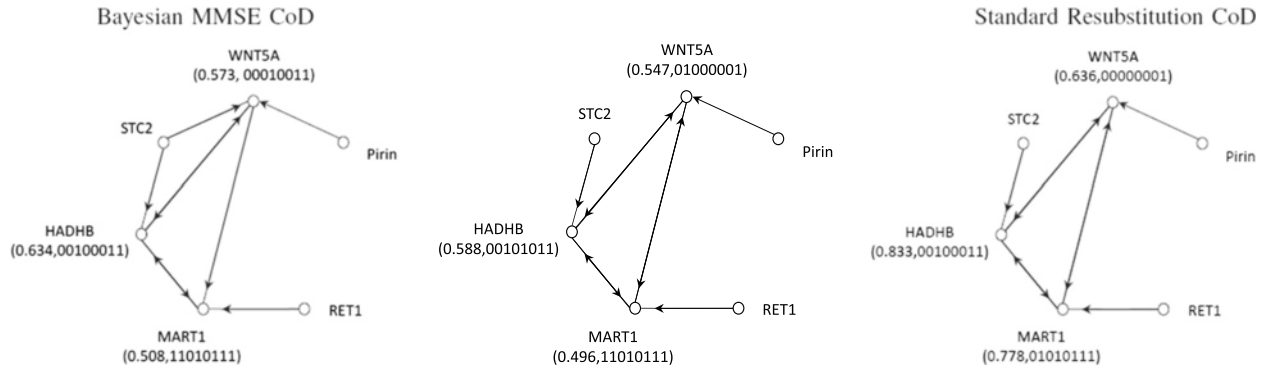

Fig. 7 Gene regulatory networks inferred using the Bayesian model selection procedure for the Bayesian MMSE CoD and OBP CoD and the standard model selection procedure using the resubstitution CoD, corresponding to the top three predicted relationships according to CoD magnitude

Appendix A: the beta distribution

If $X \sim \text{Beta}(a,b)$, where $a,b > 0$, then the probability density function of X is given by

$$f_X(u) = \frac{1}{\text{B}(a,b)} u^{a-1}(1-u)^{b-1}, \quad 0 < u < 1, \tag{35}$$

where the normalizing term $\text{B}(a,b)$ is known as the Beta function:

$$\text{B}(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}\,du. \tag{36}$$

Clearly, $B(a,b) = B(b,a)$.
For $k > -a$ and $l > -b$,

$$\begin{aligned} E\left[X^k(1-X)^l\right] &= \int_0^1 u^k(1-u)^l f(u)\,du \\ &= \frac{1}{\text{B}(a,b)}\int_0^1 u^{a+k-1}(1-u)^{b+l-1}\,du \\ &= \frac{\text{B}(a+k,b+l)}{\text{B}(a,b)}. \end{aligned} \tag{37}$$

For example, $E[X] = B(a+1,b)/B(a,b) = a/(a+b)$ (the second equality can be proved using the definition of the Beta function in terms of the Gamma function and the properties of the latter [28]). Similarly, $E[1/X] = B(a,b-1)/B(a,b) = (a+b-1)/(b-1)$, provided that $b > 1$.

The *incomplete* Beta function is defined as

$$\text{IB}(x;a,b) = \int_0^x u^{a-1}(1-u)^{b-1}\,du, \quad 0 \le x \le 1. \tag{38}$$

Notice that $\text{B}(a,b) = \text{IB}(1;a,b)$.
It is easy to verify that

$$P(X \le x) = \frac{\text{IB}(x;a,b)}{\text{B}(a,b)} \quad \text{and} \quad P(X > x) = \frac{\text{IB}(1-x;b,a)}{\text{B}(a,b)}. \tag{39}$$

Finally, for $k > -a$ and $l > -b$,

$$\begin{aligned} E\left[X^k(1-X)^l I_{X\le x}\right] &= \frac{\text{IB}(x;a+k,b+l)}{\text{B}(a,b)} I_{x<1} \\ &\quad + \frac{\text{B}(a+k,b+l)}{\text{B}(a,b)} I_{x\ge 1}, \end{aligned} \tag{40}$$

which follows easily from the definitions of the Beta density and the incomplete Beta function, and the fact that $X \in [0,1]$. In particular, if $x \ge 1$, then $E\left[X^k(1-X)^l I_{X\le x}\right] = E\left[X^k(1-X)^l\right]$.

Clearly, all the previous quantities can be computed in terms of the incomplete beta function, an expression of which is given by the next result.

Theorem 1. *If $X \sim Beta(a,b)$, then*

$$IB(x;a,b) = \sum_{i=0}^{P} r_i(a,b)\, x^{a+i}, \tag{41}$$

where $P = b-1$ and

$$r_i(a,b) = \frac{(-1)^i}{a+i}\binom{b-1}{i}, \quad i = 0,\ldots,b-1, \tag{42}$$

if b is an integer, or $P = \infty$ and

$$r_i(a,b) = \frac{(-1)^i}{a+i}\frac{(b-1)(b-2)\cdots(b-i+1)}{i!}, \quad i = 1,2,\ldots, \tag{43}$$

otherwise.

Proof. When b is a positive non-integer (that is, $\lfloor b \rfloor > 0$), we have, by using the Taylor series expansion,

$$(1-x)^{b-1} = \sum_{i=0}^{\infty}(-1)^i\binom{b-1}{i}x^i. \tag{44}$$

Note that $\lfloor b \rfloor$ denotes the largest integer that is less than b. Therefore,

$$\mathrm{IB}(k;a,b) = \int_0^k \sum_{i=0}^{\infty} (-1)^i \binom{b-1}{i} x^{a+i-1} dx. \tag{45}$$

To interchange the integration and summation in (45), we need to construct a sequence of measurable functions $g_i(x)$, $i = 0, 1, \ldots, \infty$, that satisfy the following three conditions:

(i) $\left|(-1)^i \binom{b-1}{i} x^{a+i-1}\right| \le g_i(x)$, for all k and almost all x;
(ii) $\sum_{i=0}^{\infty} g_i(x)$ converges for almost all x;
(iii) $\sum_{i=0}^{\infty} \int_0^1 g_i(x)dx < \infty$.

Let $g_i(x) = \left|\binom{b-1}{i}\right| x^{a+i-1}, i = 0, \ldots, \infty$, and obviously the condition (i) is satisfied.

For $0 \le x < 1$,

$$\sum_{i=0}^{\infty} g_i(x) = \sum_{i=0}^{\lfloor b \rfloor} g_i(x) + \sum_{i=\lfloor b \rfloor + 1}^{\infty} g_i(x), \tag{46}$$

where

$$\begin{aligned}
\sum_{i=\lfloor b \rfloor + 1}^{\infty} g_i(x) &= \sum_{i=\lfloor b \rfloor + 1}^{\infty} \left\{ \prod_{j=1}^{\lfloor b \rfloor} \left| \frac{b}{j} - 1 \right| \times \prod_{j=\lfloor b \rfloor + 1}^{i} \left| \frac{b}{j} - 1 \right| \times x^{a+i-1} \right\} \\
&= \sum_{i=\lfloor b \rfloor + 1}^{\infty} \left\{ \prod_{j=1}^{\lfloor b \rfloor} \frac{b-j}{\lfloor b \rfloor - j + 1} \times \prod_{j=\lfloor b \rfloor + 1}^{i} \left(1 - \frac{b}{j}\right) \right. \\
&\qquad \left. \times x^{a+i-1} \right\} \\
&< \sum_{i=\lfloor b \rfloor + 1}^{\infty} x^{a+i-1} \ (\text{Since } \lfloor b \rfloor + 1 > b) \\
&= \frac{x^{a+\lfloor b \rfloor}}{1-x} < \infty,
\end{aligned} \tag{47}$$

and thus the condition (ii) is satisfied.

$$\sum_{i=0}^{\infty} \int_0^1 g_i(x)dx = \frac{1}{a} + \sum_{i=1}^{\infty} \prod_{j=1}^{i} \left| \frac{b}{j} - 1 \right| \frac{1}{a+i}. \tag{48}$$

Let $\prod_{j=1}^{i} \left| \frac{b}{j} - 1 \right| \frac{1}{a+i} = z_i (i = 1, 2, \ldots, \infty)$. Since

$$\lim_{i \to \infty} i \cdot \left(\frac{z_i}{z_{i+1}} - 1 \right) = \lim_{i \to \infty} \frac{(b-i)(a+2i+1)}{(b-i+1)(a+i)} = 2 > 1, \tag{49}$$

$\sum_{i=0}^{\infty} \int_0^1 g_i(x)dx$ converges by Raabe's test [29].

Now we can interchange the integration and summation in (45), and we have

$$\begin{aligned}
\mathrm{IB}(k;a,b) &= \sum_{i=0}^{\infty} (-1)^i \binom{b-1}{i} \int_0^k x^{a+i-1} dx \\
&= \sum_{i=0}^{\infty} \frac{(-1)^i k^{a+i}}{a+i} \binom{b-1}{i}.
\end{aligned} \tag{50}$$

When b is an integer, we have $(1-x)^{b-1} = \sum_{i=0}^{b-1} (-1)^i \binom{b-1}{i} x^i$, and it is easy to show that

$$\mathrm{IB}(k;a,b) = \sum_{i=0}^{b-1} \frac{(-1)^i k^{a+i}}{a+i} \binom{b-1}{i}. \tag{51}$$

□

Notice that $B(a,b) = \sum_{i=0}^{P} r_i(a,b)$. Note also that the general case reduces to the special case if b is an integer. An equivalent expression can be derived where a appears in the binomial coefficient instead, which can then be used if a is an integer. If neither a nor b are integers, an approximation can be obtained by truncating the resulting infinite series, or by using a numerical software package.

If both a and b are integers, then $\mathrm{IB}(x;a,b)$ reduces to a polynomial in x. Otherwise, it is a simple matter to replace the finite summations by infinite series as specified in Theorem 1.

Appendix B: the Dirichlet distribution

Consider a random vector $\mathbf{X} = (X_1, \ldots, X_K)$, with $K \ge 2$, defined over the $(K-1)$-simplex

$$S_{K-1} = \left\{ (X_1, \ldots, X_K) \in R^K \mid X_i \ge 0,\ i = 1, \ldots, K,\ X_1 + \cdots + X_K = 1 \right\}.$$

If $\mathbf{X} \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_K)$, where $a_i > 0$, for $i = 1, \ldots, K$, then the probability density function of $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(x_1, \ldots, x_K) = \frac{1}{\mathrm{B}(a_1, \ldots, a_K)} \prod_{i=1}^{K} x_i^{a_i - 1}, \quad (x_1, \ldots, x_K) \in S_{K-1}, \tag{52}$$

where the normalizing term $\mathrm{B}(a_1, \ldots, a_K)$ is the multivariate generalization of the Beta function:

$$\mathrm{B}(a_1, \ldots, a_K) = \int_{S_{K-1}} \prod_{i=1}^{K} x_i^{a_i - 1} d\mathbf{x}. \tag{53}$$

The shape of the Dirichlet distributions controlled by the *concentration parameter* $\Delta = \sum_{i=1}^{K} a_i$ and the *base measure* $(a'_1, \ldots, a'_K) = (a_1/\Delta, \ldots, a_K/\Delta)$. Note that the base measure is a valid discrete probability measure. It can be shown easily that

$$E[\mathbf{X}] = (a'_1, \ldots, a'_K),$$

so that the base measure provides the "central" value around which $\mathbf{X}$ is distributed. In particular, large components in the base measure bias the distribution in their direction.

The concentration parameters, on the other hand, control the variance of the distribution around the base measure, with large values indicating smaller variance. In fact, it can be shown that [30]

$$\begin{aligned} \mathrm{Var}(X_i) &= \frac{a_i'(1-a_i')}{\Delta+1} \\ \mathrm{Cov}(X_i, X_j) &= \frac{-a_i' a_j'}{\Delta+1}, \quad \text{for} i \neq j. \end{aligned} \tag{54}$$

From the previous equations, one can see that, as Δ approaches infinity, variances converge to zero and $\mathbf{X}$ becomes equal to the base measure with probability 1; in addition, covariances also go to zero, rendering the components of $\mathbf{X}$ uncorrelated. The special case $a_i = 1$, for all $i = 1, \ldots, K$ corresponds to a uniform over S_{K-1}. This corresponds to a uniform base measure and concentration parameter $\Delta = K$. If the base measure is not uniform but $\Delta = K$, the distribution is approximately uniform. For Δ approaching zero, the distribution becomes concentrated at the boundary of the simplex.

Summing up, large Δ implies large probability density around the base measure, $\Delta = K$ implies a nearly uniform distribution, whereas Δ close to zero produces sparse sample vectors with most of the components close to zero.

The Dirichlet distribution is the multivariate generalization of the Beta distribution, in the sense that the components of a Dirichlet-distributed vector $\mathbf{X} = (X_1, \ldots, X_K)$ are Beta distributed: $X_i \sim \mathrm{Beta}(a_i, \Delta - a_i)$, for $i = 1, \ldots, K$. Notice that in the case $K = 2$ the Dirichlet distribution essentially reduces to the Beta distribution.

Competing interests
The authors declare that they have no competing interests.

About the Authors
Ting Chen received the PhD degree in electrical engineering from Texas A&M University, College Station, TX, in 2013. She worked as a postdoctoral researcher in the Department of Electrical and Computer Engineering at Texas A&M University from October 2013 to April 2014. She is now working as a bioinformatician at the Emmes Corporation, Rockville, MD. Her current research interests include the study of discrete prediction and estimation methods and their applications in genomics.
Ulisses M. Braga-Neto received the PhD degree in Electrical and Computer Engineering from The Johns Hopkins University, Baltimore, MD, in 2002. He is an Associate Professor in the Department of Electrical and Computer Engineering at Texas A&M University, College Station, TX, where he has been a faculty member since 2007. He was a post-doctoral fellow at the University of Texas M.D. Anderson Cancer Center, Houston, from 2002 to 2004, and a researcher at the Oswaldo Cruz Foundation (FIOCRUZ), in Recife, Brazil, from 2004 to 2006. Dr. Braga-Neto received an NSF CAREER Award in 2008 for his work on error estimation for pattern recognition with applications in genomic signal processing. He is a Senior Member of IEEE, was President of the Midsouth Computational Biology and Bioinformatics Society (MCBIOS) in 2010–2011, and was voted to the Texas A&M University Council of Principal Investigators in 2014. He has been associate editor for several journals and special issues. His research interests include signal processing for Boolean dynamical systems and error estimation for pattern recognition, with applications in the study of cancer and infectious diseases.

Acknowledgements
The authors acknowledge the support of the National Science Foundation, through NSF award CCF-1320884.

Author details
[1]Emmes Corporation, 401 N. Washington Street, Suite 700, Rockville, MD 20850, USA. [2]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA.

Received: 11 June 2015 Accepted: 4 December 2015
Published online: 15 January 2016

References

1. S Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor. Biol. **22**(3), 437–467 (1969)
2. S Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*. (Oxford University Press, New York, NY, 1993)
3. S Bornholdt, Boolean network models of cellular regulation: prospects and limitations. J. R. Soc. Interface. **5**(1), S85—S94 (2008)
4. R Albert, H Othmer, The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster. J. Theor. Biol. **223**(1), 1–18 (2003)
5. F Li, YLu T Long, Q Ouyang, C Tang, The yeast cell-cycle network is robustly designed. Proc. Natl. Acad. Sci. U.S.A. **101**(14), 4781–4876 (2004)
6. A Faure, A Naldi, C Chaouiya, D Thieffry, Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. Bionformatics. **22**(14), 124–131 (2006)
7. ER Dougherty, S Kim, Y Chen, Coefficient of determination in nonlinear signal processing. EURASIP J. Signal Process. **80**(10), 2219–2235 (2000)
8. S Kim, ER Dougherty, Y Chen, K Sivakumar, P Meltzer, JM Trent, M Bittner, Multivariate measurement of gene expression relationships. Genom. **67**(2), 201–209 (2000)
9. X Zhou, X Wang, ER Dougherty, Binarization of microarray data based on a mixture model. Mol. Cancer Ther. **2**(7), 679–684 (2003)
10. S Kim, ER Dougherty, ML Bittner, Y Chen, K Sivakumar, P Meltzer, JM Trent, General nonlinear framework for the analysis of gene interaction via multivariate expression arrays. J. Biomed. Opt. **5**(4), 411–424 (2000)
11. I Shmulevich, ER Dougherty, S Kim, W Zhang, Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinforma. **18**(2), 261–274 (2002)
12. D Martins, U Braga-Neto, R Hashimoto, M Bittner, ER Dougherty, Intrinsically multivariate predictive genes. IEEE J. Sel. Top. Sign. Proces. **2**(3), 424–439 (2008)
13. T Chen, UM Braga-Neto, Statistical detection of intrinsically multivariate predictive genes. IEEE/ACM Trans. Comput. Biol. Bioinform. **12**(4), 951–964 (2015)
14. T Chen, UM Braga-Neto, Exact performance of CoD estimators in discrete prediction. EURASIP J. Adv. Signal Process (2010). (Article ID 2010:487893)
15. T Chen, UM Braga-Neto, Maximum-likelihood estimation of the discrete coefficient of determination in stochastic boolean systems. IEEE Trans. Signal Process. **61**(15), 3880–3894 (2013)
16. T Chen, UM Braga-Neto, Statistical detection of Boolean regulatory relationships. IEEE/ACM Trans. Comput. Biol. Bioinform. **10**(5), 1310–1321 (2013)
17. LA Dalton, ER Dougherty, Bayesian minimum mean-square error estimation for classification error – Part I: Definition and the Bayesian mmse error estimator for discrete classification. IEEE Trans. Signal Process. **59**(1), 115–129 (2011)
18. LA Dalton, ER Dougherty, Bayesian minimum mean-square error estimation for classification error – Part II: Linear classification of gaussian models. IEEE Trans. Signal Process. **59**(1), 130–144 (2011)
19. T Chen, UM Braga-Neto, in *In Proceedings of the 2013 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS'2013)*. Optimal Bayesian MMSE estimation of the coefficient of determination for discrete prediction (TX, Houston, Nov 2013), pp. 66–69
20. LA Dalton, ER Dougherty, Optimal classifiers with minimum expected error within a Bayesian framework – Part I: Discrete and gaussian models. Pattern Recogn. **46**(5), 1301–1314 (2013)

21. LA Dalton, ER Dougherty, Optimal classifiers with minimum expected error within a Bayesian framework – Part II: Properties and performance analysis. Pattern Recogn. **46**(5), 1288–1300 (2013)
22. L Devroye, L Gyorfi, G Lugosi, *A Probabilistic Theory of Pattern Recognition*. (Springer, New York, 1996)
23. G Casella, R Berger, *Statistical Inference*, 2nd ed. (Pacific Grove, CA, Duxbury, 2002)
24. M Bittner, P Meltzer, Y Chen, Y Jiang, E Seftor, M Hendrix, M Radmacher, R Simon, Z Yakhini, A Ben-Dor, N Sampas, ER Dougherty, F Marincola, E Wang, C Gooden, J Lueders, A Glatfelter, P Pollock, J Carpten, E Gillanders, D Leja, K Dietrich, C Beaudry, M Berens, D Alberts, V Sondak, N Hayward, J Trent, Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature. **406**, 536–540 (2000)
25. S Kim, ER Dougherty, N Cao, Y Chen, M Bittner, E Suh, Can markov chain models mimic biological regulation? J. Biol. Syst. **10**, 437–458 (2002)
26. A Datta, A Choudhary, M Bittner, ER Dougherty, External control in markovian genetic regulatory networks. Mach. Learn. **52**, 169–191 (2003)
27. UM Braga-Neto, ER Dougherty, *Error Estimation for Pattern Recognition*. (Wiley, New York, 2015)
28. S Ross, *A first course in probability*, 4th ed. (Macmillan, New York, 1994)
29. G Arfken, *Mathematical Methods for Physicists*, 3rd ed. (Academic Press, Orlando, FL, 1985)
30. N Balakrishnan, V Nevzorov, *A Primer on Statistical Distributions*. (Wiley, Hoboken, NJ, 2003)