Linearized Binary Regression

Andrew S. Lan[1], Mung Chiang[2], and Christoph Studer[3]

[1]Princeton University, Princeton, NJ; andrew.lan@princeton.edu
[2]Purdue University, West Lafayette, IN; chiang@purdue.edu
[3]Cornell University, Ithaca, NY; studer@cornell.edu

Abstract—**Probit regression was first proposed by Bliss in 1934 to study mortality rates of insects. Since then, an extensive body of work has analyzed and used probit or related binary regression methods (such as logistic regression) in numerous applications and fields. This paper provides a fresh angle to such well-established binary regression methods. Concretely, we demonstrate that *linearizing* the probit model in combination with *linear estimators* performs *on par* with state-of-the-art nonlinear regression methods, such as posterior mean or maximum a-posteriori estimation, for a broad range of real-world regression problems. We derive *exact*, *closed-form*, and *nonasymptotic* expressions for the mean-squared error of our linearized estimators, which clearly separates them from nonlinear regression methods that are typically difficult to analyze. We showcase the efficacy of our methods and results for a number of synthetic and real-world datasets, which demonstrates that *linearized binary regression* finds potential use in a variety of inference, estimation, signal processing, and machine learning applications that deal with binary-valued observations or measurements.**

I. INTRODUCTION

This paper deals with the estimation of the N-dimensional vector $\mathbf{x} \in \mathbb{R}^N$ from the following measurement model:

$$\mathbf{y} = \text{sign}(\mathbf{D}\mathbf{x} + \mathbf{w}). \qquad (1)$$

Here, the vector $\mathbf{y} \in \{-1, +1\}^M$ contains M binary-valued measurements, the function $\text{sign}(z)$ operates element-wise on its argument and outputs $+1$ for $z \geq 0$ and -1 otherwise, $\mathbf{D} \in \mathbb{R}^{M \times N}$ is a given design matrix (or matrix of covariates). The noise vector $\mathbf{w} \in \mathbb{R}^M$ has i.i.d. random entries. Estimation of the vector $\mathbf{x}$ from the observation model in (1) is known as binary regression. The two most common types of binary regression are (i) *probit regression* [1] for which the noise vector $\mathbf{w}$ follows a standard normal distribution and (ii) *logistic regression* [2] for which the noise vector $\mathbf{w}$ follows a logistic distribution with unit scale parameter.

Binary regression finds widespread use in a broad range of applications and fields, including (but not limited to) image classification [3], biomedical data analysis [4], [5], economics [6], and signal processing [7], [8]. In most real-world applications, one can use either probit or logistic regression, since the noise distribution is unknown; in this paper, we focus on probit regression for reasons that we will detail in

Section II-A. In what follows, we will assume that the noise vector $\mathbf{w} \in \mathbb{R}^M$ has i.i.d. standard normal entries, and refer to (1) as the standard probit model.

A. Relevant Prior Art

1) Estimators: The two most common estimation techniques for the standard probit model in (1) are the posterior mean (PM) and maximum a-posteriori (MAP) estimators. The PM estimator computes the following conditional expectation [9]:

$$\hat{\mathbf{x}}^{\text{PM}} = \mathbb{E}_{\mathbf{x}}[\mathbf{x}|\mathbf{y}] = \int_{\mathbb{R}^N} \mathbf{x}\, p(\mathbf{x}|\mathbf{y})\mathrm{d}\mathbf{x}, \qquad (2)$$

where $p(\mathbf{x}|\mathbf{y})$ is the posterior probability of the vector $\mathbf{x}$ given the observations $\mathbf{y}$ under the model (1). The PM estimator is optimal in a sense that it minimizes the mean-squared error (MSE) defined as

$$MSE(\hat{\mathbf{x}}) = \mathbb{E}_{\mathbf{x},\mathbf{w}}\left[\|\mathbf{x} - \hat{\mathbf{x}}\|^2\right], \qquad (3)$$

and is, hence, also known as the nonlinear minimum mean-squared error (MMSE) estimator. Evaluating the integral in (2) for the probit model is difficult and hence, one typically resorts to rather slow Monte-Carlo methods [10]. By assuming that the vector $\mathbf{x}$ is multivariate Gaussian, an alternative regression technique is the MAP estimator that solves the following convex optimization problem [11]:

$$\hat{\mathbf{x}}^{\text{MAP}} = \underset{\mathbf{x} \in \mathbb{R}^N}{\arg\min} -\sum_{m=1}^{M} \log(\Phi(y_m \mathbf{d}_m^T \mathbf{x})) + \tfrac{1}{2}\mathbf{x}^T \mathbf{C}_{\mathbf{x}}^{-1}\mathbf{x}. \qquad (4)$$

Here, $\Phi(x) = \int_{-\infty}^{x}(2\pi)^{-1/2}e^{-t^2/2}\mathrm{d}t$ is the cumulative distribution function of a standard normal random variable, $\mathbf{d}_m^T$ is the mth row of the covariate matrix $\mathbf{D}$, and $\mathbf{C}_{\mathbf{x}}$ is the covariance matrix of the zero-mean multivariate Gaussian prior on the vector $\mathbf{x}$. By ignoring the prior on $\mathbf{x}$, one arrives at the well-known maximum-likelihood (ML) estimator. Compared to the PM estimator, MAP and ML estimation can be implemented efficiently either by solving a series of re-weighted least squares problems [12] or by using standard numerical methods for convex problems that scale favorably to large problem sizes [13], [14]. In contrast to such well-established nonlinear estimators, we will investigate *linear estimators* that are computationally efficient and whose performance is on par to that of the PM, MAP, and ML estimators.

2) Analytical Results: Analytical results that characterize the performance of estimation under the probit model are almost exclusively for the *asymptotic* setting, i.e., when M and/or N

tend to infinity. More specifically, Brillinger [15] has shown in 1982 that the conventional least-squares (LS) estimators for scenarios in which the design matrix $\mathbf{D}$ has i.i.d. Gaussian entries, delivers an estimate that is the same as that of the PM estimator up to a constant. More recently, Brillinger's result has been generalized by Thrampoulidis *et al.* [16] to the sparse setting, i.e., where the vector $\mathbf{x}$ has only a few nonzero entries. Other related results analyze the consistency of the ML estimator for sparse logistic regression. These results are either asymptotic [8], [17], [18] or of probabilistic nature [19]; the latter type of results bounds the MSE with high probability. In contrast to all such existing analytical results, we will provide *nonasymptotic* and *exact* expressions for the MSE that are valid for *arbitrary* and *deterministic* design matrices $\mathbf{D}$.

B. Contributions

We propose novel *linear estimators* of the form $\hat{\mathbf{x}} = \mathbf{W}\mathbf{y}$ for the probit model in (1), where $\mathbf{W} \in \mathbb{R}^{N \times M}$ are suitably-chosen estimation matrices, and provide *exact, closed-form,* and *nonasymptotic* expressions for the MSE of these estimators. Specifically, we will develop two estimators: a linear minimum mean-squared error (L-MMSE) estimator that aims at minimizing the MSE in (3) and a more efficient but less accurate least-squares (LS) estimator. Our MSE results are in stark contrast to existing performance guarantees for the MAP or PM estimators, for which a nonasymptotic performance analysis is, in general, difficult. We provide inference results on synthetic data, which suggest that the inference quality of the proposed linear estimators is on par with state-of-the-art nonlinear estimators, especially at low signal-to-noise ratio (SNR), i.e., when the quantization error is lower than the noise level. Moreover, we show using six different real-world binary regression datasets that the proposed linear estimators achieve competitive predictive performance to PM and MAP estimation at comparable or even lower complexity.

II. Linearized Probit Regression

To develop and analyze linearized inference methods for the standard probit model in (1), we will first consider the following smoothed version of the probit model:

$$\bar{\mathbf{y}} = f_\sigma(\mathbf{D}\mathbf{x} + \mathbf{w}). \tag{5}$$

We will then use these results to study the binary model (1). Here, $\bar{\mathbf{y}} \in [-1, +1]^M$, $\mathbf{x}$ is zero-mean Gaussian with known covariance $\mathbf{C_x}$, the sigmoid function is defined as $f_\sigma(z) = 2\Phi(z/\sigma) - 1$ and operates element-wise on its argument, $\sigma \in (0, \infty)$ is a smoothing parameter, and the vector $\mathbf{w}$ is assumed to be zero-mean Gaussian with known covariance $\mathbf{C_w}$ and independent of $\mathbf{x}$.[1] We emphasize that as $\sigma \to 0$, the sigmoid function $f_\sigma(z)$ corresponds to the sign function and hence, the model in (5) includes the probit model in (1) as a special case. In what follows, we assume nondegenerate covariance matrices for $\mathbf{x}$ and $\mathbf{w}$, i.e., we assume that $\mathbf{C_x}$ and $\mathbf{C_w}$ are both

---

[1]We emphasize that these are standard model assumptions in Bayesian data analysis (see, e.g., [20]) and in numerous real-world applications, such as modeling user responses to test items [21].

invertible. We next introduce two new linear estimators for this model and then, provide exact, closed-form, and nonasymptotic expressions for the associated MSEs.

A. Linear Minimum Mean-Squared Error Estimator

Our main result is as follows.

Theorem 1. *The linear minimum mean-squared error (L-MMSE) estimate for the generalized probit model in (5) is*

$$\hat{\mathbf{x}}^{\text{L-MMSE}} = \mathbf{E}^T \mathbf{C}_{\bar{\mathbf{y}}}^{-1} \bar{\mathbf{y}}, \tag{6}$$

where

$$\mathbf{E} = \left(\tfrac{2}{\pi}\right)^{1/2} \operatorname{diag}(\operatorname{diag}(\sigma^2\mathbf{I} + \mathbf{C_z})^{-1/2})\mathbf{D}\mathbf{C_x}, \tag{7}$$

$$\begin{aligned} \mathbf{C}_{\bar{\mathbf{y}}} = \tfrac{2}{\pi} \arcsin(&\operatorname{diag}(\operatorname{diag}(\sigma^2\mathbf{I} + \mathbf{C_z})^{-1/2})\mathbf{C_z} \\ &\times \operatorname{diag}(\operatorname{diag}(\sigma^2\mathbf{I} + \mathbf{C_z})^{-1/2})), \end{aligned} \tag{8}$$

and $\mathbf{C_z} = \mathbf{D}\mathbf{C_x}\mathbf{D}^T + \mathbf{C_w}$.

Remark 1. *The reason that we focus on probit regression is that under the standard probit model, the matrices $\mathbf{E}$ and $\mathbf{C}_{\bar{\mathbf{y}}}$ exhibit closed-form expressions; For logistic regression, such closed-form expressions do not exist.*

Proof. The proof consists of two steps. First, we linearize the model in (5). Then, we derive the L-MMSE estimate in (6) for the linearized model. The two steps are as follows.

Step 1 (Linearization): Let $\mathbf{z} = \mathbf{D}\mathbf{x} + \mathbf{w}$ and

$$\bar{\mathbf{y}} = f_\sigma(\mathbf{z}) = \mathbf{F}\mathbf{x} + \mathbf{e} \tag{9}$$

be a linearization of the generalized probit model in (5), where $\mathbf{F} \in \mathbb{R}^{M \times N}$ is a *linearization matrix* and $\mathbf{e} \in \mathbb{R}^M$ is a *residual error vector* that contains noise and linearization artifacts. Our goal is to perform a Bussgang-like decomposition [22], which uses the linearization matrix $\mathbf{F}$ that minimizes the ℓ_2-norm of the residual error vector $\mathbf{e}$ averaged over the signal and noise. Concretely, let $\mathbf{C_z}$ be the covariance matrix of the vector $\mathbf{z}$ and consider the optimization problem

$$\underset{\mathbf{F} \in \mathbb{R}^{M \times N}}{\operatorname{minimize}} \mathbb{E}_{\mathbf{x},\mathbf{w}}\left[\left\|\bar{\mathbf{y}} - \mathbf{F}\mathbf{x}\right\|^2\right],$$

which has a closed-form solution that is given by $\mathbf{F} = \mathbf{E}\mathbf{C}_{\mathbf{x}}^{-1}$ with $\mathbf{E} = \mathbb{E}_{\mathbf{x},\mathbf{w}}\left[\bar{\mathbf{y}}\mathbf{x}^T\right]$. It can easily be verified that for this particular choice of the linearization matrix $\mathbf{F}$, the residual error vector $\mathbf{e}$ and the signal of interest $\mathbf{x}$ are uncorrelated, i.e., we have $\mathbb{E}_{\mathbf{x},\mathbf{w}}\left[\mathbf{x}\mathbf{e}^T\right] = \mathbf{0}_{N \times M}$.

We now derive a closed-form expression for the entries of the matrix $\mathbf{E}$. Since both $\mathbf{x}$ and $\mathbf{w}$ are independent and zero-mean Gaussian, the bivariate (z_m, x_n) is jointly Gaussian for each index pair $\{m, n\}$. Moreover, we have $\mathbb{E}_t[|f_\sigma(t)|] < \infty$ and $\mathbb{E}_t[|tf_\sigma(t)|] < \infty$ if t is a zero-mean Gaussian random variable. Hence, we can use the following result that is due to Brillinger [23, Lem. 1]:

$$[\mathbf{E}]_{m,n} = \mathbb{E}_{\mathbf{x},\mathbf{w}}[\bar{y}_m x_n] = \frac{Cov(z_m, x_n)}{Var(z_m)} \mathbb{E}_{z_m}[\bar{y}_m z_m], \tag{10}$$

where $Cov(z_m, x_n) = \mathbf{d}_m^T \mathbf{c}_n$ with $\mathbf{c}_n$ being the nth column of $\mathbf{C_x}$. Since for $\sigma > 0$ the function $\bar{y}_m = f_\sigma(z_m)$ is absolutely

continuous[2], z_m is zero-mean Gaussian, and $\mathbb{E}_t[f'_\sigma(t)] < \infty$, we can invoke Stein's Lemma [24], which states that

$$\frac{\mathbb{E}_{z_m}[f_\sigma(z_m)z_m]}{Var(z_m)} = \mathbb{E}_{z_m}[f'_\sigma(z_m)], \qquad (11)$$

with $f'_\sigma(z) = \frac{d}{dz}f_\sigma(z)$. Using $f_\sigma(x) = 2\Phi(x/\sigma) - 1$, we can evaluate the right-hand side in (11) as

$$\begin{aligned}
\mathbb{E}_{z_m}[f'_\sigma(z_m)] &= 2\,\mathbb{E}_{z_m}[\Phi'(z_m/\sigma)] \\
&= \frac{2}{\sigma}\int_{-\infty}^{\infty}\mathcal{N}(z_m/\sigma; 0, 1)\,\mathcal{N}(z_m; 0, \gamma_m)dz_m \\
&= \frac{2}{\sigma}\frac{\sigma'}{\sqrt{2\pi\gamma_m}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma'}\exp\left(-\frac{z_m^2}{2\sigma'^2}\right)dz_m \\
&= \left(\frac{2}{\pi}\right)^{1/2}\frac{1}{\sqrt{\sigma^2+\gamma_m}}, \qquad (12)
\end{aligned}$$

where $\mathcal{N}(z; \mu, \sigma^2)$ denotes the probability density function of a Gaussian distribution with mean μ and variance σ^2 evaluated at z, $\gamma_m = Var(z_m) = \mathbf{d}_m^T\mathbf{C_x}\mathbf{d}_m + [\mathbf{C_w}]_{m,m}$, and $\sigma'^2 = \frac{\sigma^2\gamma_m}{\sigma^2+\gamma_m}$. Combining (10) with (11) and (12) leads to

$$[\mathbf{E}]_{m,n} = \left(\frac{2}{\pi}\right)^{1/2}\frac{\mathbf{d}_m^T\mathbf{c}_n}{\sqrt{\sigma^2+\mathbf{d}_m^T\mathbf{C_x}\mathbf{d}_m+[\mathbf{C_w}]_{m,m}}},$$

where (7) represents the entire matrix $\mathbf{E}$ in compact notation.

Step 2 (L-MMSE Estimator): We have linearized the probit model as $\bar{\mathbf{y}} = f_\sigma(\mathbf{z}) = \mathbf{Fx} + \mathbf{e}$ in (9) with $\mathbf{F} = \mathbf{EC_x^{-1}}$. We now estimate $\mathbf{x}$ from this linearization using the L-MMSE estimator. Since the residual distortion vector $\mathbf{e}$ is uncorrelated to the vector $\mathbf{x}$, the L-MMSE estimator is given by

$$\hat{\mathbf{x}}^{\text{L-MMSE}} = \mathbf{E}^T\mathbf{C}_{\bar{\mathbf{y}}}^{-1}\bar{\mathbf{y}},$$

where $\mathbf{C}_{\bar{\mathbf{y}}} = \mathbb{E}_{\mathbf{x},\mathbf{w}}[\bar{\mathbf{y}}\bar{\mathbf{y}}^T]$ is the covariance matrix of the generalized probit measurements in (5). The remaining piece is to calculate the individual entries of this matrix.

With abuse of notation, we start by deriving the necessary expressions for a general pair of correlated but zero-mean Gaussian random variables (x, y) with covariance matrix $\mathbf{C} = [C_{x,x}, C_{x,y}; C_{x,y}, C_{y,y}]$. More specifically, we are interested in computing the quantity

$$\begin{aligned}
\mathbb{E}_{x,y}[f_\sigma(x)f_\sigma(y)] &= 4\,\mathbb{E}_{x,y}[\Phi(x/\sigma)\Phi(y/\sigma)] + 1 \\
&\quad - 2\,\mathbb{E}_x[\Phi(x/\sigma)] - 2\,\mathbb{E}_y[\Phi(y/\sigma)].
\end{aligned}$$

Since

$$\begin{aligned}
\mathbb{E}_x[\Phi(x/\sigma)] &= \int_{-\infty}^{\infty}\Phi(x/\sigma)\mathcal{N}(x; 0, C_{x,x})dx \\
&= \int_0^{\infty}(\Phi(-x/\sigma)+\Phi(x/\sigma))\mathcal{N}(x; 0, C_{x,x})dx = \tfrac{1}{2},
\end{aligned}$$

we have

$$\mathbb{E}_{x,y}[f_\sigma(x)f_\sigma(y)] = 4\,\mathbb{E}_{x,y}[\Phi(x/\sigma)\Phi(y/\sigma)] - 1. \qquad (13)$$

Hence, we only need a closed-form expression for $\mathbb{E}_{x,y}[\Phi(x/\sigma)\Phi(y/\sigma)]$, which we derive using direct integration. We rewrite this expression as follows:

$$\mathbb{E}_{x,y}\left[\Phi\left(\frac{x}{\sigma}\right)\Phi\left(\frac{y}{\sigma}\right)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Phi\left(\frac{x}{\sigma}\right)\Phi\left(\frac{y}{\sigma}\right)$$

$$\times \mathcal{N}\left(\left[\begin{array}{c} x \\ y \end{array}\right]; \mathbf{0}, \left[\begin{array}{cc} C_{x,x} & C_{x,y} \\ C_{x,y} & C_{y,y} \end{array}\right]\right)dxdy$$

$$= \int_{-\infty}^{\infty}\Phi\left(\frac{x}{\sigma}\right)\Phi\left(\frac{x}{\sigma'}\right)\mathcal{N}(x; 0, C_{x,x})dx,$$

where the last equality follows from [25, Sec. 3.9] with $\sigma' = \frac{C_{x,x}}{C_{x,y}}\sqrt{\sigma^2+C_{y,y}+\frac{C_{x,y}^2}{C_{x,x}}}$. We now further simplify the above expression with the following steps:

$$\begin{aligned}
\mathbb{E}_{x,y}&\left[\Phi\left(\frac{x}{\sigma}\right)\Phi\left(\frac{y}{\sigma}\right)\right] \\
&= \int_{-\infty}^{\infty}\Phi\left(\frac{x}{\sigma}\right)\Phi\left(\frac{x}{\sigma'}\right)\frac{1}{\sqrt{2\pi C_{x,x}}}\exp\left(-\frac{x^2}{2C_{x,x}}\right)dx \\
&= \int_{-\infty}^{\infty}\Phi\left(\frac{\sqrt{C_{x,x}}}{\sigma}x\right)\Phi\left(\frac{\sqrt{C_{x,x}}}{\sigma'}x\right)\mathcal{N}(x; 0, 1)dx.
\end{aligned}$$

Using the definitions $\sigma_1 = \sigma/\sqrt{C_{x,x}}$ and $\sigma_2 = \sigma'/\sqrt{C_{x,x}}$, we can rewrite the above expression as

$$\begin{aligned}
\mathbb{E}_{x,y}\left[\Phi\left(\frac{x}{\sigma}\right)\Phi\left(\frac{y}{\sigma}\right)\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\frac{x}{\sigma_1}}\int_{-\infty}^{\frac{x}{\sigma_2}} \\
&\mathcal{N}(y; 0, 1)\mathcal{N}(z; 0, 1)dzdy\,\mathcal{N}(x; 0, 1)dx.
\end{aligned}$$

To evaluate this expression, it is key to observe that it corresponds to the cumulative probability density of a 3-dimensional normal random variable with zero mean and an identity covariance matrix on a region cut by two planes. Imagine a cuboid with edge lengths $\{1, 1/\sigma_1, 1/\sigma_2\}$. Assume $C_{x,y} > 0$ without loss of generality. The first plane has the normal vector $[1, -\sigma_1, 0]^T$, while the second plane has the normal vector $[1, 0, -\sigma_2]^T$. To find a convenient way to evaluate this integral, we need to find an appropriate change of coordinates. Define the first new coordinate x' as the intersection of the two planes, along the direction of $[1, 1/\sigma_1, 1/\sigma_2]^T$. With proper normalization, this implies $x' = \frac{\sigma_1\sigma_2 x+\sigma_2 y+\sigma_1 z}{\sqrt{\sigma_1^2\sigma_2^2+\sigma_1^2+\sigma_2^2}}$. Then, we let the second coordinate y' be orthogonal to x' and also to the first plane, i.e., orthogonal to the normal vector of the first plane, $[1, 1/\sigma_1, 1/\sigma_2]^T$. This gives $y' = \frac{\sigma_1^2 x+\sigma_1 y-\sigma_2(\sigma_1^2+1)z}{\sqrt{\sigma_1^2+1}\sqrt{\sigma_1^2\sigma_2^2+\sigma_1^2+\sigma_2^2}}$. The third coordinate is simply $z' = \frac{x-\sigma_1 y}{\sqrt{\sigma_1^2+1}}$, taken as the normal vector to the first plane. The unit vector in the second plane that is orthogonal to x' and y' is given by $v' = \frac{\sigma_2^2 x-\sigma_1(\sigma_2^2+1)y-\sigma_2 z}{\sqrt{\sigma_2^2+1}\sqrt{\sigma_1^2\sigma_2^2+\sigma_1^2+\sigma_2^2}}$. Since the new coordinates form a Cartesian system and are properly normalized, the determinant of the Jacobian is one, and the covariance matrix of the 3-dimensional normal random variable remains an identity matrix. We first integrate over x' to obtain

$$\mathbb{E}_{x,y}\left[\Phi\left(\frac{x}{\sigma}\right)\Phi\left(\frac{y}{\sigma}\right)\right] = \iint_{\mathcal{C}}\mathcal{N}(y'; 0, 1)\mathcal{N}(z'; 0, 1)dy'dz',$$

where we have used $\mathcal{C}$ to denote the space to integrate over for the variables y' and z'. Since $\mathcal{C}$ is the area between the directions of y' and v' in the 2-dimensional plane, we use polar coordinates $y' = \rho\cos\theta$ and $z' = \rho\sin\theta$ to get

$$\mathbb{E}_{x,y}\left[\Phi\left(\frac{x}{\sigma}\right)\Phi\left(\frac{y}{\sigma}\right)\right]$$

$$= \int_0^{\frac{\pi}{2}+\arcsin\left(\frac{1}{\sqrt{\sigma_1^2+1}\sqrt{\sigma_2^2+1}}\right)} \int_0^\infty \frac{1}{2\pi} e^{-\frac{\rho^2}{2}} \rho\, \mathrm{d}\rho\, \mathrm{d}\theta$$

$$= \frac{1}{4} + \frac{1}{2\pi} \arcsin\left(\frac{C_{x,y}}{\sqrt{\sigma^2+C_{x,x}}\sqrt{\sigma^2+C_{y,y}}}\right).$$

Consequently, we have

$$\mathbb{E}_{x,y}[f_\sigma(x)f_\sigma(y)] = \frac{2}{\pi} \arcsin\left(\frac{C_{x,y}}{\sqrt{\sigma^2+C_{x,x}}\sqrt{\sigma^2+C_{y,y}}}\right),$$

which allows us, in combination with (13), to express the desired covariance matrix $\mathbf{C}_{\bar{\mathbf{y}}}$ as in (8). ∎

For the L-MMSE estimator in Theorem 1, we can extract the MSE in closed form:

Lemma 2. *The MSE of the L-MMSE estimator in Theorem 1 is given by*

$$MSE(\hat{\mathbf{x}}^{\text{L-MMSE}}) = \mathrm{tr}(\mathbf{C}_\mathbf{x} - \mathbf{E}^T \mathbf{C}_{\bar{\mathbf{y}}}^{-1} \mathbf{E}).$$

Proof. The proof follows from the MSE definition in (3) and the facts that $\mathbf{F} = \mathbf{E}\mathbf{C}_\mathbf{x}^{-1}$ and the two vectors $\mathbf{x}$ and $\mathbf{e}$ are uncorrelated for the L-MMSE estimator in (6). ∎

By letting the parameter $\sigma \to 0$ in (5), we can use Theorem 1 and Lemma 2 to obtain the following corollary for the standard probit model in (1). This result agrees with a recent result in wireless communications [26].

Corollary 3. *The L-MMSE estimate for the standard probit model in (1) is $\hat{\mathbf{x}}^{\text{L-MMSE}} = \mathbf{E}^T \mathbf{C}_\mathbf{y}^{-1} \mathbf{y}$, where*

$$\mathbf{E} = \left(\frac{2}{\pi}\right)^{1/2} \mathrm{diag}(\mathrm{diag}(\mathbf{C}_\mathbf{z})^{-1/2})\mathbf{D}\mathbf{C}_\mathbf{x},$$
$$\mathbf{C}_\mathbf{y} = \frac{2}{\pi} \arcsin(\mathrm{diag}(\mathrm{diag}(\mathbf{C}_\mathbf{z})^{-1/2})\mathbf{C}_\mathbf{z}$$
$$\times \, \mathrm{diag}(\mathrm{diag}(\mathbf{C}_\mathbf{z})^{-1/2})),$$

and $\mathbf{C}_\mathbf{z} = \mathbf{D}\mathbf{C}_\mathbf{x}\mathbf{D}^T + \mathbf{C}_\mathbf{w}$. The associated MSE is given by $MSE(\hat{\mathbf{x}}^{\text{L-MMSE}}) = \mathrm{tr}(\mathbf{C}_\mathbf{x} - \mathbf{E}^T \mathbf{C}_\mathbf{y}^{-1} \mathbf{E}).$

B. Least Squares (LS) Estimator

The L-MMSE estimator as in (6) requires the computation of $\mathbf{C}_{\bar{\mathbf{y}}}$ followed by a matrix inversion. For large-scale problems, one can avoid the matrix inversion by first solving $\mathbf{y} = \mathbf{C}_{\bar{\mathbf{y}}}\mathbf{q}$ for $\mathbf{q}$ using conjugate gradients [13], followed by calculating $\hat{\mathbf{x}}^{\text{L-MMSE}} = \mathbf{E}^T\mathbf{q}$. Hence, the complexity of L-MMSE estimation is comparable to that of MAP estimation. Computation of $\mathbf{C}_{\bar{\mathbf{y}}}$, however, cannot be avoided entirely.

Fortunately, there exists a simpler linear estimator that avoids computation of $\mathbf{C}_{\bar{\mathbf{y}}}$ altogether, which we call the least-squares (LS) estimator. Concretely, let $M \geq N$ and consider the linearization in (9), which is $\bar{\mathbf{y}} = f_\sigma(\mathbf{z}) = \mathbf{E}\mathbf{C}_\mathbf{x}^{-1}\mathbf{x} + \mathbf{e}$. By ignoring the residual error vector $\mathbf{e}$ and by assuming that the columns of $\mathbf{E}$ are linearly independent, we can simply invert the matrix $\mathbf{E}\mathbf{C}_\mathbf{x}^{-1}$, which yields the LS estimate

$$\hat{\mathbf{x}}^{\text{LS}} = \mathbf{C}_\mathbf{x}\mathbf{E}^+\bar{\mathbf{y}}, \qquad (14)$$

where $\mathbf{E}^+ = (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T$ is the left pseudo-inverse of $\mathbf{E}$. Again, one can use conjugate gradients to implement (14). In contrast to the L-MMSE estimator, the LS estimator does not require knowledge of $\mathbf{C}_{\bar{\mathbf{y}}}$, which makes it more efficient yet slightly less accurate (see the experimental results section for a comparison). As for the L-MMSE estimator, we have a closed-form expression for the MSE of the LS estimator.

Lemma 4. *Assume that $\mathbf{E}^+$ exists. Then, the MSE of the LS estimator in (14) is given by*

$$MSE(\hat{\mathbf{x}}^{\text{LS}}) = \mathrm{tr}(\mathbf{C}_\mathbf{x}\mathbf{E}^+\mathbf{C}_{\bar{\mathbf{y}}}(\mathbf{E}^+)^T\mathbf{C}_\mathbf{x} - \mathbf{C}_\mathbf{x}).$$

Proof. The proof follows from the MSE definition (3), and the facts that $\mathbb{E}[\bar{\mathbf{y}}\mathbf{x}^T] = \mathbf{E}$ and $\mathbf{E}^+\mathbf{E} = \mathbf{I}$. ∎

III. NUMERICAL RESULTS

We now experimentally demonstrate the efficacy of the proposed linear estimators.

A. Experiments with Synthetic Data

We first compare the MSE of our estimators to that of the nonlinear MAP and PM estimators using synthetic data.

1) Experimental Setup: We set the dimensions of $\mathbf{x}$ to $N \in \{5, 20\}$ and the number of measurements to $M \in \{10, 50, 200\}$. We first generate a single random matrix $\mathbf{D}$ of size $M \times N$ with i.i.d. standard normal entries, and normalize each row to have unit ℓ_2-norm. Then, we generate the entries of $\mathbf{x}$ from a multivariate normal distribution with zero mean and covariance matrix $\mathbf{C}_\mathbf{x} = \sigma_x^2\mathbf{I}$. The entries of the noise vector $\mathbf{w}$ are i.i.d. zero-mean Gaussian with variance σ_w^2. The vector $\mathbf{y}$ is generated using the standard probit model in (1). We sweep the SNR defined as $SNR = \sigma_x^2/\sigma_w^2$ by changing the noise variance σ_w^2. For the PM estimator, we use a standard Gibbs sampling procedure [10]; we use the mean of the generated samples over $50{,}000$ iterations as the PM estimate after a burn-in phase of $20{,}000$ iterations. For the MAP estimator, we use an accelerated gradient-descent procedure [14], [27] to solve (4) up to machine precision with a maximum number of $20{,}000$ iterations. We repeat all experiments for 100 trials and report the empirical MSE.

2) Results and Discussion: Fig. 1 shows the MSE of the L-MMSE, LS, MAP, and PM estimators. We do not show LS for $M = 10$ and $N = 20$ as it does not exist if $M < N$. We see that at low SNR ($SNR \leq 0\,\text{dB}$), the L-MMSE, MAP, and PM estimators achieve a similar MSE. Hence, linearizing the probit model does not entail a noticeable performance degradation if the measurements are noisy. At higher SNR, the performance of the different estimators varies. For a small number of measurements, the performance of L-MMSE estimation is superior to MAP estimation. For a large number of measurements (e.g., $M = 200$), the MSE of L-MMSE estimation is slightly higher than that of MAP estimation for some SNR values. We note that the MSE performance of MAP degrades with increasing SNR, and we observe that there is an optimal SNR level for MAP estimation; this observation is in line with those reported in [28] for 1-bit matrix completion using ML-type estimators. Per design, PM estimation achieves the lowest MSE for all configurations, but is notoriously slow. We conclude that linearized probit regression entails a negligible
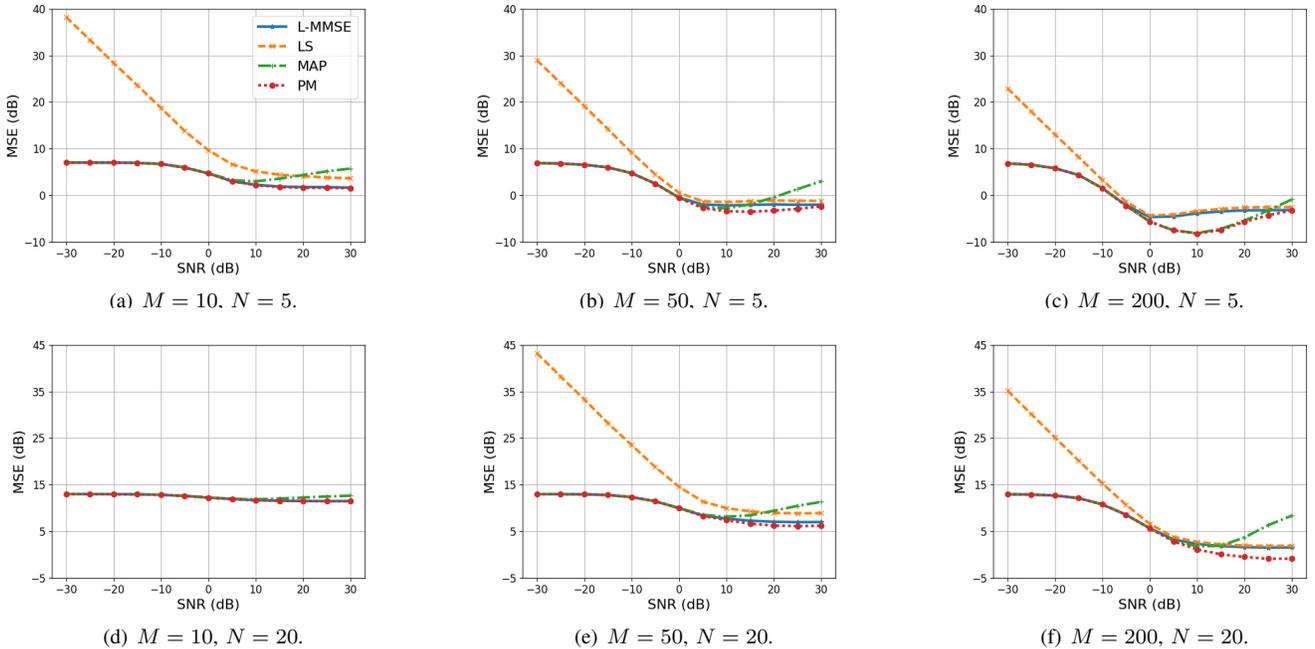
Fig. 1. Mean squared error (MSE) versus signal-to-noise ratio (SNR) for various problem sizes $M \in \{10, 50, 200\}$ and dimensions $N \in \{5, 20\}$. We see that for most parameter settings, the MSE of the proposed L-MMSE estimator is comparable to that of the optimal PM estimator; MAP estimation and LS estimation do not work as well at high and low SNR, respectively.

MSE performance loss compared to PM estimation, for a wide range of parameter settings.

B. Experiments with Real-World Data

We now validate the performance of the proposed linearized estimators using a variety of real-world datasets. Since the noise model in real-world datasets is generally unknown, we also consider the performance of MAP estimation using the logistic noise model (indicated by "Logit-MAP").

1) Datasets: We use a range of standard binary regression datasets in this experiment. These datasets include (i) "Admissions", which consists of binary-valued graduate school admission outcomes and features of the applicants, with $M = 400$ and $N = 3$, (ii) "Lowbwt", which consists of low child birthweight indicators and features of their parents, with $M = 109$ and $N = 10$, (iii) "Polypharm", which consists of whether an adult takes more than one type of prescription and their features, with $M = 3,499$ and $N = 15$, (iv) "Myopia", which consists of myopia test outcomes for adults and their personal features, with $M = 575$ and $N = 11$, (v) "Uis", which consists of treatment outcomes for AIDS patients and their personal features, with $M = 618$ and $N = 15$, and (vi) "SAheart", which consists of whether a person has heart disease and their features, with $M = 462$ and $N = 9$. The first five datasets are taken from [29] and the last one is from [12].

2) Experimental Setup: We evaluate the prediction quality of the L-MMSE, LS, MAP, PM, and Logit-MAP estimators using five-fold cross validation. We randomly divide the entire dataset into five nonoverlapping subsets, use four folds of the data as the training set and the other fold as the test set. We use the training set to estimate $\mathbf{x}$, and use it to predict the binary-valued outcomes on the test set. For all experiments, we fix $\sigma_w^2 = 1$. Since the variance σ_x^2 serves as a regularization parameter for $\mathbf{x}$, we select an optimal value of σ_x^2 using grid search on a separate validation set. To assess the performance of these estimators, we deploy the two most common metrics that characterize prediction quality: prediction accuracy (ACC) and area under the receiver operating characteristic curve (AUC) [30]. Both metrics take values in $[0, 1]$ and larger values indicate better prediction performance.

3) Results and Discussion: Tables I and II show the mean and standard deviation of the performance of each estimator on both metrics across 20 random training/test partitions of the datasets. We observe that the performance of L-MMSE, MAP, PM, and Logit-MAP are virtually indistinguishable on most datasets. LS estimation is, with a few exceptions, slightly worse than all the other estimators.

We find it surprising that *linearized probit regression* performs equally well as significantly more sophisticated nonlinear estimators on a broad range of real-world datasets. We also note that the proposed linearized estimators can be implemented efficiently and scale well to large datasets, which is in stark contrast to the PM estimator.

IV. CONCLUSIONS

We have shown that linearizing the well-known probit regression model in combination with linear estimators is able to achieve comparable estimation performance to nonlinear methods such as MAP and PM estimators for binary regression

TABLE I

MEAN AND STANDARD DEVIATION OF PREDICTION QUALITY IN TERMS OF PREDICTION ACCURACY (ACC) FOR THE L-MMSE, LS, MAP, PM, AND LOGIT-MAP ESTIMATORS ON VARIOUS REAL-WORLD DATASETS.

	L-MMSE	LS	MAP	PM	Logit-MAP
Admissions	0.691 ± 0.036	0.691 ± 0.038	0.692 ± 0.037	$\mathbf{0.693 \pm 0.036}$	0.692 ± 0.040
Lowbwt	0.703 ± 0.070	0.707 ± 0.071	$\mathbf{0.715 \pm 0.064}$	0.713 ± 0.067	0.712 ± 0.070
Polypharm	0.779 ± 0.015	0.777 ± 0.016	$\mathbf{0.780 \pm 0.015}$	$\mathbf{0.780 \pm 0.015}$	$\mathbf{0.780 \pm 0.015}$
Myopia	0.882 ± 0.025	0.879 ± 0.024	$\mathbf{0.890 \pm 0.022}$	0.890 ± 0.023	0.890 ± 0.022
Uis	0.745 ± 0.010	$\mathbf{0.746 \pm 0.041}$	0.736 ± 0.041	0.737 ± 0.041	0.736 ± 0.041
SAheart	0.727 ± 0.042	0.726 ± 0.044	0.728 ± 0.042	$\mathbf{0.730 \pm 0.042}$	0.729 ± 0.042

TABLE II

PREDICTION QUALITY IN TERMS OF THE AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE (AUC) FOR THE L-MMSE, LS, MAP, PM, AND LOGIT-MAP ESTIMATORS ON VARIOUS REAL-WORLD DATASETS.

	L-MMSE	LS	MAP	PM	Logit-MAP
Admissions	$\mathbf{0.675 \pm 0.056}$	0.672 ± 0.054	0.674 ± 0.056	0.674 ± 0.056	$\mathbf{0.675 \pm 0.056}$
Lowbwt	$\mathbf{0.716 \pm 0.076}$	0.712 ± 0.081	$\mathbf{0.716 \pm 0.076}$	0.713 ± 0.076	0.711 ± 0.080
Polypharm	0.728 ± 0.022	0.728 ± 0.022	0.728 ± 0.022	0.728 ± 0.022	$\mathbf{0.729 \pm 0.022}$
Myopia	0.864 ± 0.038	0.862 ± 0.040	$\mathbf{0.873 \pm 0.036}$	0.873 ± 0.036	0.873 ± 0.035
Uis	0.632 ± 0.052	0.632 ± 0.051	$\mathbf{0.634 \pm 0.052}$	0.634 ± 0.051	0.633 ± 0.052
SAheart	0.769 ± 0.049	0.768 ± 0.049	0.770 ± 0.049	0.771 ± 0.049	$\mathbf{0.771 \pm 0.049}$

problems. Our linear estimators enable an exact, closed-form, and nonasymptotic MSE analysis, which is in stark contrast to existing analytical results for the MAP and PM estimators. We hence believe that the proposed linear estimators have the potential to be used in a variety of machine learning or statistics applications that deal with binary-valued observations.

REFERENCES

[1] C. Bliss, "The method of probits," *Science*, vol. 79, no. 2037, pp. 38–39, Jan. 1934.

[2] D. Cox, "The regression analysis of binary sequences," *J. R. Stat. Soc. B Met.*, vol. 20, no. 2, pp. 215–242, Mar. 1958.

[3] Y. Qian, M. Ye, and J. Zhou, "Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 2276–2291, Sep. 2013.

[4] G. Cawley and N. Talbot, "Gene selection in cancer classification using sparse logistic regression with Bayesian regularization," *Bioinformatics*, vol. 22, no. 19, pp. 2348–2355, July 2006.

[5] J. Zhu and T. Hastie, "Classification of gene microarrays by penalized logistic regression," *Biostatistics*, vol. 5, no. 3, pp. 427–443, July 2004.

[6] T. Amemiya, "Qualitative response models: A survey," *J. Econ. Lit.*, vol. 19, no. 4, pp. 1483–1536, Dec. 1981.

[7] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 957–968, July 2005.

[8] Y. Plan and R. Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *IEEE Trans. Inf. Th.*, vol. 59, no. 1, pp. 482–494, Sep. 2013.

[9] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Springer Science & Business Media, 2013.

[10] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *J. Am. Stat. Assoc.*, vol. 88, no. 422, pp. 669–679, June 1993.

[11] C. Bliss, "The calculation of the dosage-mortality curve," *Ann. Appl. Biol.*, vol. 22, no. 1, pp. 134–167, Feb. 1935.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2010.

[13] J. Nocedal and S. Wright, *Numerical Optimization*. Springer, 2006.

[14] T. Goldstein, C. Studer, and R. G. Baraniuk, "A field guide to forward-backward splitting with a FASTA implementation," *arXiv preprint: 1411.3406*, Nov. 2014.

[15] D. Brillinger, *A Festschrift for Erich L. Lehmann*, ser. Wadsworth Statistiscs/Probability Series. Chapman & Hall/CRC, 1982, ch. "A generalized linear model with "Gaussian" regressor variables", pp. 97–114.

[16] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Lasso with non-linear measurements is equivalent to one with linear measurements," in *Proc. Adv. Neural Info. Proc. Syst.*, Dec. 2015, pp. 3420–3428.

[17] F. Bunea, "Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization," *Electron. J. Stat.*, vol. 2, pp. 1153–1194, 2008.

[18] P. Ravikumar, M. Wainwright, and J. Lafferty, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Ann. Stat.*, vol. 38, no. 3, pp. 1287–1319, June 2010.

[19] F. Bach, "Self-concordant analysis for logistic regression," *Electron. J. Stat.*, vol. 4, pp. 384–414, 2010.

[20] P. D. Hoff, *A First Course in Bayesian Statistical Methods*. Springer, 2009.

[21] F. Lord, *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum Associates, 1980.

[22] J. Bussgang, "Cross-correlation function of amplitude-distorted gaussian signals," MIT, Res. Lab. Elec. Tech. Rep. 216, Mar. 1952.

[23] D. Brillinger, "The identification of a particular nonlinear time series system," *Biometrika*, vol. 64, no. 3, pp. 509–515, Dec. 1977.

[24] C. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Stat.*, vol. 9, no. 6, pp. 1135–1151, Nov. 1981.

[25] C. Rasmussen and C. Williams, *Gaussian Process for Machine Learning*. MIT Press, 2006.

[26] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Sig. Proc.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.

[27] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk, "Sparse factor analysis for learning and content analytics," *J. Mach. Learn. Res.*, vol. 15, pp. 1959–2008, June 2014.

[28] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Inf. Infer.*, vol. 3, no. 3, pp. 189–223, Sep. 2014.

[29] D. Hosmer Jr, S. Lemeshow, and R. Sturdivant, *Applied Logistic Regression*. John Wiley & Sons, 2013.

[30] H. Jin and C. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005.