**48th IEEE Conference on Decision and Control, Shanghai, China, December 2009**

# Mixed Linear System Estimation and Identification

A. Zymnis S. Boyd D. Gorinevsky

***Abstract*— We consider a mixed linear system model, with both continuous and discrete inputs and outputs, described by a coefficient matrix and a set of noise variances. When the discrete inputs and outputs are absent, the model reduces to the usual noise-corrupted linear system. With discrete inputs only, the model has been used in fault estimation, and with discrete outputs only, the system reduces to a probit model. We consider two fundamental problems: Estimating the model input, given the model parameters and the model output; and identifying the model parameters, given a training set of input-output pairs. The estimation problem leads to a mixed Boolean-convex optimization problem, which can be solved exactly when the number of discrete variables is small enough. In other cases the estimation problem can be solved approximately, by solving a convex relaxation, rounding, and possibly, carrying out a local optimization step. The identification problem is convex and so can be exactly solved. Adding $\ell_1$ regularization to the identification problem allows us to trade off model fit and model parsimony. We illustrate the identification and estimation methods with a numerical example.**

## I. Introduction

### A. System model

In this paper we introduce a model with multiple continuous (*i.e.*, real valued) and discrete (*i.e.*, Boolean) inputs and outputs. The continuous outputs are linearly related to the continuous and discrete inputs and are corrupted by zero mean Gaussian noise of known variance. The discrete outputs indicate whether a linear combination of the continuous and discrete inputs, corrupted by zero mean Gaussian noise, is above or below a known threshold. As such, the model that we consider is a hybrid generalized linear model (GLM) [1], [2], [3].

The model has the form

$$\begin{aligned} y &= A^{\rm cc}x + A^{\rm dc}d + b^{\rm c} + v_c, & (1)\\ z &= \mathbf{pos}(A^{\rm cd}x + A^{\rm dd}d + b^{\rm d} + v_d), & (2)\end{aligned}$$

where

- $x \in \mathbf{R}^{n_c}$ is the continuous input,
- $d \in \{0,1\}^{n_d}$ is the discrete input,
- $y \in \mathbf{R}^{m_c}$, is the continuous measurement or output,
- $z \in \{0,1\}^{m_d}$ is the discrete measurement or output,
- $v_c \in \mathbf{R}^{m_c}$ is the continuous noise term, and
- $v_d \in \mathbf{R}^{m_d}$ is the discrete noise term.

The function $\mathbf{pos} : \mathbf{R}^{m_d} \to \{0,1\}^{m_d}$ is defined by

$$\mathbf{pos}(u)_i = \begin{cases} 1, & u_i > 0 \\ 0, & u_i \leq 0. \end{cases}$$

Note that for any diagonal positive matrix $D$, we have $\mathbf{pos}(Du) = \mathbf{pos}(u)$.

The noises are Gaussian, with all components independent, with

$$\begin{aligned} (v_c)_i &\sim \mathcal{N}(0, \sigma_i^2), \quad i = 1, \ldots, n_c, \\ (v_d)_i &\sim \mathcal{N}(0, 1), \quad i = 1, \ldots, n_d. \end{aligned}$$

(We can assume the discrete noise components have unit variance without loss of generality, using a positive diagonal scaling.)

The model is defined by the matrices $A^{\rm cc}$, $A^{\rm dc}$, $A^{\rm cd}$, and $A^{\rm dd}$, the intercept terms $b^{\rm c}$ and $b^{\rm d}$, and the continuous noise variances $\sigma_i^2$, $i = 1, \ldots, n_c$.

The model (1)–(2) includes several well known and widely used special cases. For $m_d = 0$ and $n_d = 0$ we have a simple linear model with additive Gaussian noise. For $n_c = 0$ and $n_d = 1$, we obtain a probit model [1]. We mention several applications in §I-D.

In this paper we address two basic problems associated with this model: estimation and identification.

### B. Estimation

We first look at the problem of estimating the model inputs $x$ and $d$, given one or more output samples. The prior distribution on the inputs is specified by a density $p(x)$ for $x$, which we assume is log-concave, and the probability $p_j$ that $d_j = 1$. (We assume that $x$ and all $d_j$ are independent.) We will see that the maximum a posteriori (MAP) estimate of $(x, d)$ is the solution of a mixed Boolean convex problem. If $n_d$ is small enough, this problem can be solved exactly, by exhaustive enumeration of the possible values of $d$, or by a branch-and-bound or other global optimization method. For other cases, we propose to solve the optimization

This material is based upon work supported by the Focus Center for Circuit & System Solutions (C2S2), by JPL award I291856, by the Precourt Institute on Energy Efficiency, by Army award W911NF-07-1-0029, by NSF award 0529426, by DARPA award N66001-06-C-2021, by NASA award NNX07AEIIA, by AFOSR award FA9550-06-1-0514, and by AFOSR award FA9550-06-1-0312.

The authors are with the Department of Electrical Engineering, Stanford University (e-mail: {azymnis, boyd, gorin}@stanford.edu).

problem approximately, by solving a convex relaxation, and rounding the result, possibly followed by a local optimization step. We refer to the resulting estimate as the RMAP ('relaxed MAP') estimate. Numerical simulation suggests that the estimation performance of the RMAP estimator is quite similar to the MAP estimate; unlike the MAP estimate, however, it is computationally tractable even for very large problems [4].

### *C. Identification*

We then look at the dual problem of fitting a model of the form (1)–(2), *i.e.*, determining values for the model parameters, given a given set of (training) data samples

$$(x^{(1)}, d^{(1)}, y^{(1)}, z^{(1)}), \quad \ldots, \quad (x^{(K)}, d^{(K)}, y^{(K)}, z^{(K)}).$$

We show that the associated log-likelihood function is a concave function of the model parameters, so maximum likelihood (ML) model fitting reduces to solving a convex optimization problem. To obtain a parsimonious model, *i.e.*, one in which many of the entries of the parameter matrices are zero, we propose $\ell_1$-regularized ML fitting (which is also a convex problem). By varying a regularization parameter, we can trade off model fit and model parsimony [5], [6], [7], [8].

### *D. Prior and related work*

The model that we consider is in essence a hybrid generalized linear model (GLM). These models have been extensively used for explanatory modelling. Some good references on GLMs are [1], [2], [3]. There is a considerable amount of research that deals with special cases of our problem. For example, when $m_c = 0$ and $n_c = 0$, we get a model which is very similar to a digraph model commonly used in diagnostics, *e.g.*, see [9], [10]. The formulation in this paper is an extension of the earlier work of the authors [4], [11], [12]. In [4] we considered a special case of the current problem, in the context of fault identification. The paper [11] considers a dynamic system with continuous and discrete states and continuous outputs. The upcoming paper [12] considers a special case of sparse parametric inputs and discrete outputs, where the goal is to compute the sparsity pattern of the inputs.

## II. Estimation

In this section we consider the problem of estimating the most probable values of $(x, d)$, given measurements $(y, z)$. We first derive the log posterior probability of $(x, d)$ given $(y, z)$ and show that it is jointly concave in $x$ and $d$ (with $d$ relaxed to take values in $[0, 1]$). Thus the problem of estimating the maximum a posteriori (MAP) estimate of $x$ and $d$ is a mixed integer convex problem, *i.e.*, a convex optimization problem with the additional constraint that some of the variables take values in $\{0, 1\}$. We then present an efficient heuristic for solving this problem approximately, based on a convex relaxation of the combinatorial MAP problem. Our analysis follows closely the previous work of the authors in fault detection [4].

The method that we present is readily extended to the case when we have multiple measurements $(y^{(1)}, z^{(1)}), \ldots, (y^{(M)}, z^{(M)})$ for the same input $(x, d)$: We just have to stack all these measurements and augment the system model equations appropriately.

### *A. Maximum a posteriori estimation*

*a) Log posterior:* From Bayes' rule the posterior probability (density) of $x$ and $d$ given $y$ and $z$ is

$$p(x, d|y, z) \propto \prod_{i=1}^{m_c} p(y_i|x, d) \prod_{i=1}^{m_d} p(z_i|x, d) p(d) p(x).$$

Taking logarithms, we have

$$\log p(x, d|y, z) = l(x, d) + C,$$

where $C$ is a constant and

$$l(x, d) = l_{\mathrm{mc}}(x, d) + l_{\mathrm{md}}(x, d) + l_{\mathrm{pd}}(d) + l_{\mathrm{pc}}(x), \quad (3)$$

with the terms described below. The posterior contribution due to the continuous measurements is

$$l_{\mathrm{mc}}(x, d) = -\sum_{i=1}^{m_c} \frac{1}{2\sigma_i^2} (y - A^{\mathrm{cc}} x - A^{\mathrm{dc}} d - b^{\mathrm{c}})_i^2.$$

The posterior contribution due to the discrete measurements is

$$l_{\mathrm{md}}(x, d) = \sum_{i=1}^{m_d} z_i \log \Phi(A^{\mathrm{cd}} x + A^{\mathrm{dd}} d + b^{\mathrm{d}})_i + \sum_{i=1}^{m_d} (1 - z_i) \log \Phi(-A^{\mathrm{cd}} x - A^{\mathrm{dd}} d - b^{\mathrm{d}})_i, \quad (4)$$

where $\Phi$ is the cumulative distribution of a standard Gaussian. The discrete variable prior term is

$$l_{\mathrm{pd}}(d) = \lambda^T d,$$

where $\lambda_j = \log(p_j/(1 - p_j))$. The continuous variable prior term is

$$l_{\mathrm{pc}}(x) = \log p(x).$$

The log posterior (3) is jointly concave in $x$ and $d$, when $d$ is relaxed to take values in $[0, 1]^{n_d}$. Indeed, each of the terms is concave in $x$ and $d$: $l_{\mathrm{mc}}$ is a concave quadratic in $x$ and $d$, $l_{\mathrm{pd}}$ is a linear function of $d$, and $\log p(x)$ is concave by assumption. Concavity of $l_{\mathrm{md}}$ follows from log-concavity of $\Phi$; see, *e.g.*, [13, §3.5.2]. Figure 1 shows a plot of $\log \Phi(x)$ versus $x$ for $x$ ranging between $-5$ and $5$.

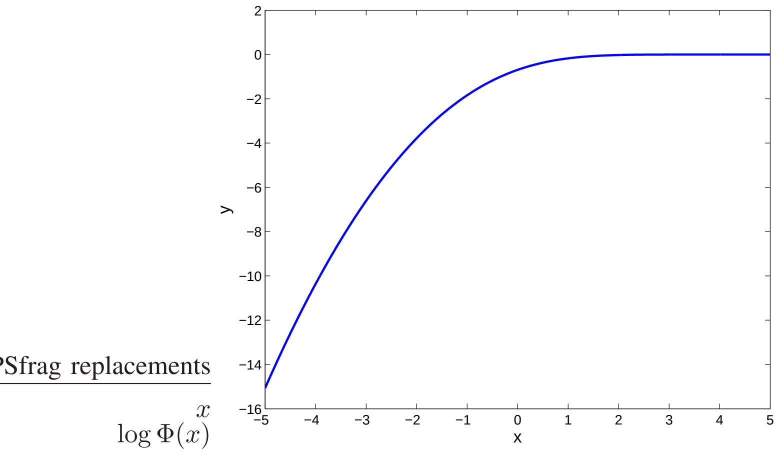

**Fig. 1:** Plot of $\log \Phi(x)$ versus $x$.

*b) MAP estimation:* The problem of estimating the most probable input variables $x$ and $d$, given the outputs $y$ and $z$, can be cast as the following optimization problem:

$$\begin{array}{ll} \text{maximize} & l(x,d) \\ \text{subject to} & d \in \{0,1\}^{n_d}, \end{array} \qquad (5)$$

with variables $x \in \mathbf{R}^{n_c}$ and $d \in \{0,1\}^{n_d}$. This is a mixed integer convex problem. One straightforward method for solving it is to enumerate all $2^{n_d}$ possible values for $d$, and to find the optimum $x$ in each case, which is tractable, since for fixed $d$ the problem is convex in $x$ [13]. This approach is not practical for $n_d$ larger than around 15 or so, or smaller, if the other dimensions are large. A branch and bound method, or other global optimization technique, can be used to (possibly) speed up computation of the globally optimal solution [14], [15]. But the worst case complexity of any method that computes the global solution is exponential in $n_d$. For this reason we need to consider heuristics for solving problem (5) approximately.

### *B. Relaxed MAP estimation*

In this section we describe a heuristic for approximately solving the MAP problem (5). Our heuristic is based on replacing the hard constraints $d_j \in \{0,1\}$ with soft constraints $d_j \in [0,1]$. This results in a convex optimization problem that we can solve efficiently. We follow this by rounding and (possibly) a simple local optimization method to improve our estimate.

*c) Linear relaxation:* We relax problem (5) to the following optimization problem

$$\begin{array}{ll} \text{maximize} & l(x,d) \\ \text{subject to} & 0 \le d \le 1, \end{array} \qquad (6)$$

with variables $x \in \mathbf{R}^{n_c}$ and $d \in \mathbf{R}^{n_d}$. This is a convex optimization problem, since it involves maximizing a concave function over a convex set. We can efficiently solve this problem in many ways, *e.g.*, via interior-point methods [16], [13]. The complexity of such methods can be shown to be cubic in $n_c + n_d$ (assuming a fixed number of iterations of an interior-point method). Since the feasible set for the relaxed MAP problem (6) contains the feasible set for the MAP problem (5), the optimal value of the relaxed MAP problem, which we denote $l_{\mathrm{ub}}$, gives an upper bound on the optimal value of the MAP problem.

Let $(x^\star, d^\star)$ be an optimal point for the relaxed MAP problem (6), so we have $l_{\mathrm{ub}} = l(x^\star, d^\star) \ge l(x^\star, d)$ for any Boolean $d$. If $d^\star$ is also Boolean, *i.e.*, $d_j^\star \in \{0,1\}$ for all $j$, we conclude that $d^\star$ is in fact optimal for the MAP problem. In other words: When a solution to the relaxed MAP problem turns out to have Boolean entries, it is optimal for the MAP problem. In general, of course, this does not happen; at least some values of $d_j^\star$ will lie between 0 and 1.

*d) Rounding:* Let $(x^\star, d^\star)$ denote the optimal point of the problem (6). We refer to $d^\star$ as a *soft decision*, since its components can be strictly between 0 and 1. The next step is to round the soft decision $d^\star$ to obtain a valid Boolean solution (or *hard decision*) for $d$. Let $\theta \in (0,1)$ and set

$$\hat{d} = \mathbf{pos}(d^\star - \theta).$$

To create $\hat{d}$, we simply round all entries of $d_j^\star$ smaller than the threshold $\theta$ to zero. Thus $\theta$ is a threshold for guessing that a discrete input variable is 1, based on the relaxed MAP solution $d^\star$. As $\theta$ varies from 0 to 1, this method generates up to $n$ different estimates $\hat{d}$, as each entry in $d$ falls below the threshold. We can efficiently find them all by sorting the entries of $d^\star$, and setting the values of $\hat{d}_j$ to one in the order of increasing $d_j^\star$.

We evaluate the log-posterior for each of these (or a subset) by solving the optimization problem

$$\text{maximize} \quad l(x, \hat{d}), \qquad (7)$$

with variables $x \in \mathbf{R}^{n_c}$. This is an unconstrained convex problem that can also be solved efficiently, again with cubic complexity in $n_c$. The RMAP continuous variable estimate $\hat{x}$ is obtained as the minimizer of (7) corresponding to the best obtained estimate $\hat{d}$.

*e) Local optimization:* Further improvement in our estimate can sometimes be obtained by a local optimization method. We describe here the simplest possible such method. We initialize $\hat{d}$ as the one which results in the largest value of $l(x,d)$ after rounding. We then cycle through $j = 1, \ldots, n$, at step $j$ replacing $\hat{d}_j$ with $1 - \hat{d}_j$. If this leads to an increase in the optimal value of problem (7), we accept the change and continue. If (as usually is the case) flipping the $j$th bit results in a decrease in $l$, we go on to the next index. We continue until we have rejected changes in all entries in $\hat{d}$. (At this point we can be sure that $\hat{d}$ is 1-OPT, which means that no change in one entry will improve the loss function.) Numerical experiments show that this local optimization method often has no effect, which means that the rounded solution is 1-OPT. In some cases, however, it can lead to a modest increase in $l$.

*f) Performance of RMAP:* The performance of our estimate of $(x, d)$ should be judged by (for example) the mean-square error in estimating $x$, and the probability of making errors in estimating $d$ (which could be further broken down into false positive and false negative error rates). Numerical examples show that RMAP has very similar performance as MAP, but has the advantage of tractability. This can be partially explained as follows. When the estimation problem is 'hard', for example, when the noise levels are high, no estimation method (and in particular, neither MAP nor RMAP) can do a good job at estimating $x$ and $d$. When the estimation problem is 'easy', for example, when the noise levels are low, even simple estimation methods (including RMAP) can do a good job at estimating $x$ and $d$. So it is only problems in between 'hard' and 'easy' where we could possibly see a significant difference in estimation performance between MAP and RMAP. In this region, however, we observe from numerical experimens that MAP and RMAP achieve very similar performance.

## III. Identification

In §II we addressed the problem of estimating $(x, d)$ given measurements $(y, z)$ when the system model is known. In this section we look at the dual problem of fitting a model of the form (1)–(2) to given data, assuming that the continuous noise variances are known. We show that the log likelihood of the model parameters given the measurements is a concave function, so we can solve the maximum likelihood (ML) problem efficiently.

We first observe that since all measurements are independent of each other, we can separately identify the model parameters that correspond to each measurement. The complexity of the resulting ML identification technique is thus linear in the total number of measurements.

Finally, we present a simple technique, variously known as compressed sensing [5], [6], [7], the Lasso [17], [18], sparse signal recovery [19], and basis pursuit [20] that can be used to identify parsimonious models that fit the data well. This involves using the $\ell_1$-norm of the parameter vector of a given linear model as a surrogate for the model sparsity.

### A. Continuous parameter identification

Suppose that we are given samples of the form $(x^{(j)}, d^{(j)}, y^{(j)})$ for $j = 1, \ldots, K$. Let $a_i^{\rm cc}$, $a_i^{\rm dc}$ denote the $i$th row of $A^{\rm cc}$, $A^{\rm dc}$ respectively. Since the continuous noise terms $v_c$ are Gaussian, the ML estimate of $(a_i^{\rm cc}, a_i^{\rm dc}, b_i^{\rm c})$ given the data, is the one that maximizes

$$l_{\rm c}(a_i^{\rm cc}, a_i^{\rm dc}, b_i^{\rm c}) = -\sum_{j=1}^{K} (y_i^{(j)} - {a_i^{\rm cc}}^T x^{(j)} - {a_i^{\rm dc}}^T d^{(j)} - b_i^{\rm c})^2.$$

We can evaluate the maximum likelihood estimates of these model parameters efficiently (via least squares), if $\sigma_i$ is known. In order to estimate a parsimonious model, as well as $\sigma_i$, we propose solving the following problem

$$\text{maximize} \quad l_{\rm c}(a_i^{\rm cc}, a_i^{\rm dc}, b_i^{\rm c}) - \mu_i^{\rm c}(\|a_i^{\rm cc}\|_1 + \|a_i^{\rm dc}\|_1), \tag{8}$$

which is a convex optimization problem, for a fixed parameter $\mu_i^{\rm c} > 0$. Having solved problem (8), we can then estimate the noise variance $\sigma_i^2$ as the variance of the measurement residuals $(y_i^{(j)} - {a_i^{\rm cc}}^T x^{(j)} - {a_i^{\rm dc}}^T d^{(j)} - b_i^{\rm c})$.

### B. Discrete parameter identification

Now suppose that we are given samples of the form $(x^{(j)}, d^{(j)}, z^{(j)})$ for $j = 1, \ldots, K$. Let $a_i^{\rm cd}$, $a_i^{\rm dd}$ denote the $i$th row of $A^{\rm cd}$, $A^{\rm dd}$ respectively. The log likelihood of $(a_i^{\rm cc}, a_i^{\rm dc}, b_i^{\rm c})$ given the data is

$$\begin{aligned} & l_{\rm d}(a_i^{\rm cd}, a_i^{\rm dd}, b_i^{\rm d}) = \\ & \quad \textstyle\sum_{j=1}^{K} \Big( z_i^{(j)} \log \Phi({a_i^{\rm cd}}^T x^{(j)} + {a_i^{\rm dd}}^T d^{(j)} + b_i^{\rm d}) \\ & \quad\quad + (1 - z_i^{(j)}) \log \Phi(-{a_i^{\rm cd}}^T x^{(j)} - {a_i^{\rm dd}}^T d^{(j)} - b_i^{\rm d}) \Big), \end{aligned}$$

which is a concave function of $(a_i^{\rm cd}, a_i^{\rm dd}, b_i^{\rm d})$. We can thus efficiently evaluate the ML estimates of the discrete model parameters efficiently, for example using Newton's method. This is equivalent to solving a probit regression problem. In order to estimate a parsimonious model we propose solving the following problem

$$\text{maximize} \quad l_{\rm d}(a_i^{\rm cd}, a_i^{\rm dd}, b_i^{\rm d}) - \mu_i^{\rm d}(\|a_i^{\rm cd}\|_1 + \|a_i^{\rm dd}\|_1), \tag{9}$$

which is a convex optimization problem, for a fixed parameter $\mu_i^{\rm d} > 0$.

### C. Computational complexity

Problems (8) and (9) can be solved efficiently in a variety of methods such as interior-point methods (in a way similar to [21], [8]) and first order methods, *e.g.*, [22]. In any case the complexity of solving this problem is cubic in $n_c + n_d$ and linear in $K$. Thus the overall complexity of $\ell_1$-regularized ML system identification is $O(K(m_c + m_d)(n_c + n_d)^3)$.

### D. Choice of regularization parameter

The regularization parameter $\mu_i^c$ and $\mu_i^d$ control the tradeoff between data fit (as measured by $l(A)$) and model sparsity (as measured by the $\ell_1$ norm). In order to keep our modelling procedure as flexible as possible, we use a different regularization parameter for each continuous and discrete sensor.

We use cross validation to choose each of the regularization parameters $\mu_i^c$ and $\mu_i^d$. We divide our available data into a training set and a test set. For each continuous sensor $i = 1, \ldots, m_c$ and for each value of $\mu_i^c$ we solve problem (8) for the training set and measure the average square residual of measurement $i$ on the test set. We then choose as $\mu_i^c$ the value of $\mu$ that gives the smallest residual. We repeat this process for the discrete sensors, where instead of square residual we use the average error rate in the discrete sensor measurement.

This is the simplest possible way of fitting the regularization parameters. There are various other methods, such as minimizing the Akaike information criterion (AIC), or minimizing the Bayesian information criterion (BIC). For a detailed description of these methods see [3, §7].

## IV. Numerical example

In this section we present the results of applying our estimation and identification methods to a small artificially generated example. Specifically, we consider a system with $n_c = n_d = 10$ and $m_c = 18$ and $m_d = 16$. We draw the elements of all system matrices randomly such that each of the matrices $A^{\rm cc}$, $A^{\rm dc}$, $A^{\rm cd}$, or $A^{\rm dd}$ are $10\%$ sparse. We draw the nonzero entries of the $A^{\rm cc}$ and $A^{\rm dc}$ matrices from a $\mathcal{N}(0,1)$ distribution, and the nonzero entries of the $A^{\rm cd}$ and $A^{\rm dd}$ matrices from a $\mathcal{N}(0,1)$ distribution. The entries of $b^{\rm c}$ and $b^{\rm d}$ are drawn from a $\mathcal{N}(0,1)$ and a $\mathcal{N}(0,100)$ distribution respectively.

We set $\sigma_i = 0.01$ for all $i$. Each element of input $x$ is generated according to a $\mathcal{N}(0,1)$ prior. The prior probability of $d_j$ being equal to 1 is 0.2 for all $j$.

We generate three sets of 500 samples: a training set, which we use to fit our estimated model, a validation set, which we use to select the best value of the regularization parameter $\mu$ for each sensor, and a test set, on which we judge the accuracy of our resulting model.

We look at 30 candidate values of $\mu$, logarithmically spaced in the range $(10^{-2}, 10^3)$. For each value of $\mu$, we use the method described in §III to fit a mixed linear model to the given training set. For each continuous sensor, we choose the value of $\mu$ which achieves the minimum relative root mean square error $r_y$ on the validation set, defined as

$$r_y = \frac{\|y - \hat{y}\|_2}{\|y\|_2},$$

where $\hat{y}$ is the output predicted by the estimated model, given the true input. For each discrete sensor we choose the value of $\mu$ that achieves the minimum average error rate on $e_z$ on the predicted value of $z$ for the validation set. As an example, figure 2 shows the plot of $e_z$ versus $\mu$ for discrete sensor 5.

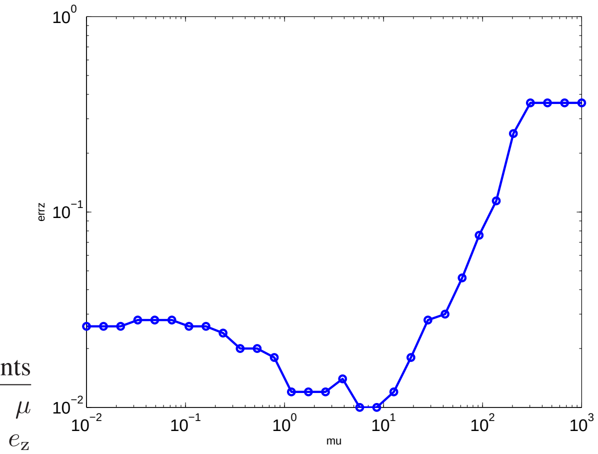

**Fig. 2:** Plot of $e_z$ as a function of $\mu$ on the validation set for discrete sensor 5.

We then use this model estimate to predict the inputs $(x, d)$ for the outputs $(y, z)$ for the test set, using the relaxed MAP method described in §II. We compute the error rate in our estimate of $d$ (which we call $e_d$), as well as the average relative root-mean-square (RMS) error between our estimate and the true value of $x$, defined as

$$r_x = \frac{\|x - \hat{x}\|_2}{\|x\|_2}.$$

Our estimated model yields an error rate $e_d$ of about $2.2 \times 10^{-3}$ on this data and a relative RMS error in $x$ of about $4.7 \times 10^{-2}$. In contrast, using the true model for estimation, yields an error rate $e_d$ of about $2.0 \times 10^{-3}$ and a relative RMS error in $x$ of about $4.1 \times 10^{-2}$. We thus see that our system identification method does a reasonable job at fitting such a model to the given data.

The modelling and estimation performance of our estimated model is shown in tables I and II respectively. We judge the modelling performance of the model, based on how well this model can predict the outputs from the inputs. As we can see from table I, the model does a good job at this task, given that the value of $r_y$ and $e_z$ for the test set are of the same order of magnitude

|  | Train | Validation | Test |
|---|---|---|---|
| $r_y$ | 0.0056 | 0.0058 | 0.0057 |
| $e_z$ | 0.0063 | 0.0080 | 0.0110 |

**TABLE I:** Modelling performance of estimated model.

|  | Train | Validation | Test |
|---|---|---|---|
| $r_x$ | 0.0500 | 0.0470 | 0.0460 |
| $e_d$ | 0.0052 | 0.0048 | 0.0022 |

**TABLE II:** Estimation performance of estimated model.

as the same values for the training and validation sets. Furthermore, from table II, we see that the model also generalizes well for estimation. The values of $r_x$ and $e_d$ for the test set are of the same order as the ones for the training and validation sets.

## V. Conclusions

We have introduced a class of mixed linear systems that includes the standard linear model and the probit model as special cases. We have presented a simple heuristic for estimating the input given the output of such a system based on a convex relaxation of a combinatorial MAP problem. We have also presented a simple heuristic that uses $\ell_1$-regularized ML to identify such a model given a set of data points. We have shown that this method performs well in practice through a numerical example.

## References

[1] P. McCullagh and J. Nelder, *Generalized linear models*. Chapman & Hall, 1989.

[2] T. Hastie and R. Tibshirani, "Generalized additive models," 1990.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2001.

[4] A. Zymnis, S. Boyd, and D. Gorinevsky, "Relaxed maximum a posteriori fault identification," *Signal Processing*, vol. 89, no. 6, pp. 989–999, 2009.

[5] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[6] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2005.

[7] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[8] K. Koh, S.-J. Kim, and S. Boyd, "An interior point method for large-scale $\ell_1$-regularized logistic regression," *Journal of Machine Learning Research*, vol. 8, pp. 1519–1555, July 2007.

[9] I. Sacks, "Digraph matrix analysis," *IEEE transactions on reliability*, vol. 34, no. 5, pp. 437–446, 1985.

[10] S. Deb, K. Pattipati, V. Raghavan, M. Shakeri, and R. Shrestha, "Multi-signal flow graphs: a novel approach for system testability analysis and fault diagnosis," *IEEE Aerospace and Electronic Systems Magazine*, vol. 10, no. 5, pp. 14–25, 1995.

[11] A. Zymnis, S. Boyd, and D. Gorinevsky, "Mixed state estimation for a linear Gaussian Markov model," in *47th IEEE Conference on Decision and Control*, 2008, pp. 3219–3226.

[12] A. Zymnis, S. Boyd, and E. Candès, "Compressed sensing with quantized measurements," 2009, in preparation.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[14] E. Lawler and D. Wood, "Branch-and-bound methods: A survey," *Operations Research*, vol. 14, pp. 699–719, 1966.

[15] R. Moore, "Global optimization to prescribed accuracy," *Computers and Mathematics with Applications*, vol. 21, no. 6–7, pp. 25–39, 1991.

[16] J. Nocedal and S. Wright, *Numerical Optimization*. Springer, 1999.

[17] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[18] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of statistics*, pp. 407–451, 2004.

[19] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[20] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, pp. 129–159, 2001.

[21] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale $\ell_1$-regularized least squares," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007.

[22] E. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for $\ell_1$-minimization: Methodology and convergence," *SIAM Journal on Optimization*, vol. 19, pp. 1107–1130, 2008.