IEEE Trans Med Imaging. 2010 Sep;29(9):1573-611. Epub 2010 Mar 18.
DOI 10.1109/TMI.2010.2044512

A statistical framework for optimal design matrix generation with application to fMRI

Gautam Pendse*, Richard Baumgartner, Adam Schwarz, Alexandre Coimbra, David Borsook and Lino Becerra

***Abstract*—The general linear model (GLM) is a well established tool for analyzing functional magnetic resonance imaging (fMRI) data. Most fMRI analyses via GLM proceed in a massively univariate fashion where the same design matrix is used for analyzing data from each voxel. A major limitation of this approach is the locally varying nature of signals of interest as well as associated confounds. This local variability results in a potentially large bias and uncontrolled increase in variance for the contrast of interest. The main contributions of this paper are two fold (1) We develop a statistical framework that enables estimation of an optimal design matrix while explicitly controlling the bias variance decomposition over a set of potential design matrices and (2) We develop and validate a numerical algorithm for computing optimal design matrices for general fMRI data sets. The implications of this framework include the ability to match optimally the magnitude of underlying signals to their true magnitudes while also matching the "null" signals to zero size thereby optimizing both the sensitivity and specificity of signal detection. By enabling the capture of multiple profiles of interest using a single contrast (as opposed to an F-test) in a way that optimizes for both bias and variance enables the passing of first level parameter estimates and their variances to the higher level for group analysis which is not possible using F-tests. We demonstrate the application of this approach to *in vivo* pharmacological fMRI data capturing the acute response to a drug infusion, to task-evoked, block design fMRI and to the estimation of a haemodynamic response function (HRF) in event-related fMRI. Although developed with motivation from fMRI, our framework is quite general and has potentially wide applicability to a variety of disciplines.**

***Index Terms*—Optimization, design matrix, bias variance tradeoff, residual, general linear model (GLM), functional magnetic resonance imaging (fMRI).**

I. INTRODUCTION

GENERAL linear models (GLMs) with Gaussian noise are very popular tools for fMRI model-based analyses [1]. The design matrix (DM) for GLM analysis is usually based on the stimulus paradigm used during the experiment. With each column or explanatory variable (EV) of the DM is associated a parameter estimate (PE) measuring the strength of that EV in the overall model fit. The investigator defines linear contrasts of interest to extract meaningful values reflecting aspects of the brain's response to the applied paradigm. Since fMRI data is composed of thousands of measured timeseries across different points or voxels in the brain ($\sim$30000 voxels is typical), GLM based analysis for fMRI proceeds in a massively univariate way, meaning that the same DM is used to analyze all voxels.

One very attractive property of the PEs estimated using GLM is that they are unbiased and of minimum variance if the DM is correctly specified (Gauss-Markov theorem) [2]. However, the exact mechanism underlying fMRI signal generation is extremely complex and in fMRI data the 'true' signal of interest is often superimposed with various artifactual signals due to physiology, motion and possible scanner effects. Moreover, the true temporal profile of the signal of interest may not be constant across brain regions or subjects; that is, there might be a range of temporal response profiles induced by the same paradigm. Thus, the assumption of a correctly specified model using a single DM for all voxels often does not hold in real fMRI data. In view of this fact, when using a GLM framework to analyze such data, one must have a good handle on the bias and variance of imperfect PEs calculated using the mis-specified DM. This is an extremely important point that cannot be ignored in fMRI analysis especially because the implications of Gauss-Markov theorem do not hold for mis-specified DMs. If the bias and variance introduced in the PEs at the first (individual subject) level are uncontrolled then misleading results can be obtained when generalizing to a group of subjects.

Small modeling misspecifications of certain types can be corrected to some extent using simple approaches, for example, by adding the temporal derivative of the main EV to the DM to capture small temporal shifts [3], [4], [5], [6], [7].

*G. V. Pendse (corresponding author) is with the Imaging and Analysis Group (IMAG), McLean Hospital, Harvard Medical School, Belmont MA 02478 USA and with the Imaging Consortium for Drug Development (ICD), McLean Hospital, Belmont MA 02478 USA. (e-mail: gpendse@mclean.harvard.edu)

R. Baumgartner is with Biometrics Research, Merck Research Laboratories, Rahway NJ 07065 USA and with the Imaging Consortium for Drug Development (ICD), McLean Hospital, Belmont MA 02478 USA.

A. J. Schwarz is with the Translational Imaging Group, Lilly Research Laboratories, Indianapolis IN 46225 USA and with the Imaging Consortium for Drug Development (ICD), McLean Hospital, Belmont MA 02478 USA.

A. Coimbra is with the Imaging Department, Merck Research Laboratories, Westpoint PA 19486 USA and with the Imaging Consortium for Drug Development (ICD), McLean Hospital, Belmont MA 02478 USA.

D. Borsook is with the Imaging and Analysis Group (IMAG), McLean Hospital, Harvard Medical School, Belmont MA 02478 USA, the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown MA 02129 USA, Harvard Medical School, Boston MA 02115 USA and with the Imaging Consortium for Drug Development (ICD), McLean Hospital, Belmont MA 02478 USA.

L. Becerra is with the Imaging and Analysis Group (IMAG), McLean Hospital, Harvard Medical School, Belmont MA 02478 USA, the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown MA 02129 USA, Harvard Medical School, Boston MA 02115 USA and with the Imaging Consortium for Drug Development (ICD), McLean Hospital, Belmont MA 02478 USA.

Manuscript received September 28, 2009; revised January 7, 2010. The Associate Editor responsible for coordinating the review of this paper was Dr. Hu Xiaoping.

However, these additional EVs can still result in uncontrolled bias and variance of the resulting PEs. The use of a signed F-like test statistic using the main EV and its temporal derivative has been proposed by [5] to correct for amplitude bias for small temporal delays. This statistic is constructed using the squared PEs and the sign of the main PE. As discussed in [8], these approaches work only for small delays (~ 1 sec). What about delays more than 1 sec? What is the general solution?

Basis function approaches result in more flexibility by allowing arbitrary response shapes to be matched via appropriately specified regressors [9]. It is possible to achieve a low variance fit to the data using these basis functions but it is difficult to define a meaningful contrast of interest that captures the underlying signal amplitude. In addition, the PEs estimated are again not controlled for bias or variance.

What about general variations in response profiles (e.g. variations in shape)? How to generate the optimal DM in a non-parametric way to minimize the bias and variance of the parameters of interest? In this article, we wish to derive a general theoretical basis that enables computation of optimal DM's for general GLM analyses as well as provide an algorithm that is practical to implement for practitioners.

This paper is organized as follows:

1) First, we develop an algorithmic framework to derive automatically both a meaningful contrast and a DM simultaneously or, given a specified contrast derive a DM suitable for use at all voxels that models the set of all potential DM's in an optimal way, capturing a wide range of potential signals of interest while controlling for both the bias and variance of the signal amplitude measure. This explicit optimization will automatically optimize both the sensitivity and specificity of detecting signals of interest in the data. Please see sections II and III.
2) Second, we apply the framework to specific case studies arising from both pharmacological challenge, task-evoked fMRI experiments and to the estimation of a haemodynamic response function (HRF) in event-related fMRI. We apply the optimal design matrices to the real fMRI data and demonstrate the more robust detection of fMRI responses *in vivo*. Please see sections IV and V.
3) Supplementary material such as proofs and derivations can be found in the appendix section IX.

II. Theoretical development

In this section, we will try to answer the following questions:

1) What happens when data generated by the unknown true model is analyzed using a proposed model? Answering this question in section II-B will allow us to measure the quality of any proposed design matrix or basis set.
2) How can we use the performance measure from II-B to set up an optimization problem whose solution gives us the optimal model or design matrix? This development is given in sections II-C and II-D.

A. Notation

We will indicate the dimension of each variable introduced in the text to clarify if the variable is a scalar, vector or a matrix. Uppercase letters such as X, Z etc. will be used to denote matrices. Vectors and scalars will be denoted by lower case letters and greek letters such as y, β, γ etc. Whether a variable is a scalar or a vector should be clear from context. I_n will denote the $n \times n$ identity matrix. A bold face zero ($\mathbf{0}$) will be used to denote a matrix or vector of all zeros the size of which should be clear from context. Estimate of a variable will be denoted by putting a hat ($\hat{}$) on top of the symbol used for the variable. For instance, an estimate of variable γ will be denoted by $\hat{\gamma}$. The trace of a matrix A will be denoted by $tr(A)$.

B. Performance measure for a single design matrix

To start let us assume that the true model generating the data is

$$y = X\beta + \varepsilon \tag{1}$$

where $X \in \mathbf{R}^{n \times q}$, $\beta \in \mathbf{R}^q$, $y \in \mathbf{R}^n$ and $\varepsilon \sim N(0, \sigma^2 I_n)$. [If the noise is non-white then the same discussion in this section applies after an initial pre-whitening step. A complication is that there might be interaction between pre-whitening and model misspecification.]

Unfortunately we do not know what X is and so we use a design matrix $Z \in \mathbf{R}^{n \times p}$ for analyzing the data generated by the above model. In other words, the assumed model is:

$$y = Z\gamma + \varepsilon_1 \tag{2}$$

where $\gamma \in \mathbf{R}^p$ and $\varepsilon_1 \sim N(0, \sigma_1^2 I_n)$. Suppose that the matrix Z has full column rank then the usual GLM estimates based on the assumed model are:

$$\hat{\gamma} = (Z^T Z)^{-1} Z^T y \tag{3}$$

and

$$\hat{\text{Cov}}(\hat{\gamma}) = \hat{\sigma}_1^2 \, (Z^T Z)^{-1} \tag{4}$$

where

$$\hat{\sigma}_1^2 = \frac{(y - Z\hat{\gamma})^T (y - Z\hat{\gamma})}{n - p} \tag{5}$$

For a contrast of interest $c_X \in \mathbf{R}^q$ for the true model 1, let $c_Z \in \mathbf{R}^p$ be the corresponding contrast of interest in the proposed model 2. When the data y comes from the true model 1 then it can be shown (see Proposition 9.1) that the following holds:

$$E(\hat{\gamma}) = (Z^T Z)^{-1} Z^T X\beta \tag{6}$$

$$(n - p)\frac{\hat{\sigma}_1^2}{\sigma^2} \sim \chi^2(n - p, \Delta) \tag{7}$$

where $\chi^2(n-p, \Delta)$ is a non-central χ^2 distribution with degrees of freedom $(n-p)$ and non-centrality parameter Δ (see appendix IX-A for a definition of non-central χ^2 distribution). The non-centrality parameter Δ is given by:

$$\Delta = \frac{\beta^T X^T P_Z X\beta}{\sigma^2} \tag{8}$$

where

$$P_Z = I_n - Z(Z^TZ)^{-1}Z^T \tag{9}$$

When $Z = X$ in 6, 7, 8 and 9 then we recover the usual GLM quantities. Ideally we would like the misspecified model to perform well, i.e, we would like $c_Z^T\hat{\gamma}$ to be as close to $c_X^T\beta$ as possible and at the same time have as small *estimated* variance as possible. This bias-variance tradeoff is captured in the estimated mean squared error (EMSE) function:

$$\begin{aligned}\text{EMSE} &= \hat{\text{Var}}(c_Z^T\hat{\gamma}) + \left(E(c_Z^T\hat{\gamma}) - c_X^T\beta\right)^2 \\ &= \hat{\sigma}_1^2\, c_Z^T(Z^TZ)^{-1}c_Z + \left(E(c_Z^T\hat{\gamma}) - c_X^T\beta\right)^2\end{aligned} \tag{10}$$

The first term on the right hand side of 10 is the estimated variance while the second term is the squared bias of the estimator $c_Z^T\hat{\gamma}$ for $c_X^T\beta$. Notice that we use the estimated variance $\hat{\text{Var}}(c_Z^T\hat{\gamma}) = \hat{\sigma}_1^2 c_Z^T(Z^TZ)^{-1}c_Z$ instead of the true variance $\text{Var}(c_Z^T\hat{\gamma}) = \sigma^2 c_Z^T(Z^TZ)^{-1}c_Z$ in 10 because we assume that the true noise variance σ^2 is unknown and therefore must be replaced by an estimate $\hat{\sigma}_1^2$. Note that this function captures simultaneously not only the bias and variance of the contrast of interest but also the full model residual (via $\hat{\sigma}_1$). Next, we normalize EMSE from 10 simply by division using σ^2.

$$F = \frac{\text{EMSE}}{\sigma^2} = \frac{\hat{\sigma}_1^2}{\sigma^2}\, c_Z^T(Z^TZ)^{-1}c_Z + \frac{1}{\sigma^2}\left(E(c_Z^T\hat{\gamma}) - c_X^T\beta\right)^2 \tag{11}$$

Using 6 and 7 the expected value of the above under the true model can be written as:

$$\begin{aligned}E(F) &= \left(1 + \frac{\beta^TX^TP_ZX\beta}{(n-p)\sigma^2}\right) c_Z^T(Z^TZ)^{-1}c_Z \\ &+ \frac{1}{\sigma^2}\left(c_Z^T(Z^TZ)^{-1}Z^TX\beta - c_X^T\beta\right)^2\end{aligned} \tag{12}$$

Normalization by σ^2 in 11 renders everything on the right hand side of the expression for $E(F)$ only a function of the signal to noise ratio $\frac{\beta}{\sigma}$ in the data. This is useful since $E(F)$ now becomes invariant to particular values of β and σ as long as the ratio $\frac{\beta}{\sigma}$ remains the same. Function $E(F)$ from 12 which captures the fundamental bias-variance tradeoff when using the estimator $c_Z^T\hat{\gamma}$ for $c_X^T\beta$ will be used as a performance measure for a given design matrix or basis set Z and contrast c_Z. This performance measure will be used in the next section where we will combine $E(F)$ across multiple potential design matrices to create a composite objective function for optimization.

C. Calculating optimal design matrices

In this section we will set up the optimization problem that will enable us to compute optimal design matrices for arbitrary data sets. Ideally we would like the PEs from our optimal design matrix to have nice properties such as a low bias, low variance as well as a "low residual" overall model fit. It can be seen that 11 will be small when a candidate DM satisfies these ideals as compared to another that does not. Hence 11 is a joint performance measure that captures all attributes of interest in one function for a given design matrix X. How do we generalize this concept to enable good performance of the optimal DM over a range of candidate DMs? Suppose our data is expected to be described by one of the m design matrices $X_1, X_2, \ldots, X_m$. Matrix X_i is of size $n \times p_i$, where p_i is the number of regressors in X_i. Suppose noisy data is generated from X_i at SNR β_i/σ_i where $\beta_i \in \mathbf{R}^{p_i}$ and $\sigma_i \in \mathbf{R}$. Suppose that the contrast of interest for X_i is $c_{X_i} \in \mathbf{R}^{p_i}$. Expanding 12 we get:

$$\begin{aligned}&f(Z, c_Z; X_i; \frac{\beta_i}{\sigma_i}; c_{X_i}) = \\ &\qquad c_Z^T(Z^TZ)^{-1}c_Z\left[1 + tr\left(P_Z\frac{X_i\beta_i\beta_i^TX_i^T}{(n-p)\sigma_i^2}\right)\right] \\ &\qquad + \frac{1}{\sigma_i^2}c_Z^T(Z^TZ)^{-1}Z^TX_i\beta_i\beta_i^TX_i^TZ(Z^TZ)^{-1}c_Z \\ &\qquad - \frac{2}{\sigma_i^2}c_Z^T(Z^TZ)^{-1}Z^TX_i\beta_i\beta_i^Tc_{X_i} + (c_{X_i}^T\beta_i/\sigma_i)^2\end{aligned} \tag{13}$$

Suppose weights $w_1, w_2, \ldots, w_m$ measure the frequency of occurance of each DM X_i in the data such that higher values of w_i indicate a higher frequency and $\sum_{i=1}^m w_i = 1$. The objective function of interest is the mean performance measure over all design matrices. Hence, we define the following composite objective function (leaving off the multiplier $\frac{1}{\sum_{i=1}^m w_i}$):

$$G(Z, c_Z) = \sum_{i=1}^m w_i f(Z, c_Z; X_i; \frac{\beta_i}{\sigma_i}; c_{X_i}) \tag{14}$$

Define the quantities:

$$\Sigma = \begin{pmatrix} \frac{1}{n-p} & 0 & \cdots \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{n-p} \end{pmatrix} \tag{15}$$

$$H = \left(\frac{\sqrt{w_1}X_1\beta_1}{\sigma_1}, \ldots, \frac{\sqrt{w_i}X_i\beta_i}{\sigma_i}, \ldots, \frac{\sqrt{w_m}X_m\beta_m}{\sigma_m}\right) \tag{16}$$

and

$$\ell = \left(\frac{\sqrt{w_1}c_{X_1}^T\beta_1}{\sigma_1}, \ldots, \frac{\sqrt{w_i}c_{X_i}^T\beta_i}{\sigma_i}, \ldots, \frac{\sqrt{w_m}c_{X_m}^T\beta_m}{\sigma_m}\right) \tag{17}$$

With these definitions the composite objective 14 can be written as:

$$\begin{aligned}G(Z, c_Z) &= c_Z^T(Z^TZ)^{-1}c_Z\left[\sum_{i=1}^m w_i + tr\left(P_ZH\Sigma H^T\right)\right] \\ &+ c_Z^T(Z^TZ)^{-1}Z^THH^TZ(Z^TZ)^{-1}c_Z \\ &- 2c_Z^T(Z^TZ)^{-1}Z^TH\ell \\ &+ \sum_{i=1}^m w_i(c_{X_i}^T\beta_i/\sigma_i)^2\end{aligned} \tag{18}$$

In general, one can put constraints on the columns of Z (e.g., fixing certain columns) such as:

$$ZA = B, \tag{19}$$

where $A \in R^{p\times q}$ and $B \in R^{n\times q}$ are fixed matrices. Similar constraints can be imposed on the contrast vector:

$$Cc_Z = d \tag{20}$$

where $C \in R^{r \times p}$ is a fixed matrix and $d \in R^r$ is a fixed vector.

Our goal is to minimize the composite objective function $G(Z, c_Z)$ that measures the weighted bias/variance decomposition over all potential DMs in the data. Hence, the complete optimization problem is written out as:

$$\hat{Z}, \hat{c_Z} = \text{arg min }_{Z,c_Z} G(Z, c_Z) \tag{21}$$

$$\text{s.t. } ZA = B \tag{22}$$

$$\text{s.t. } Cc_Z = d \tag{23}$$

$$\text{s.t. } \text{rank}(Z) = p \tag{24}$$

The last constraint above simply fixes the rank of Z or the number of independent columns in Z.

D. Local control of bias and variance

It is straightforward to extend the concepts developed above to attain a local control of bias-variance decomposition i.e., to weigh the contribution of bias and variance terms to the overall performance measure for each DM X_i. The first step is modifying 13 to accomodate user defined bias/variance weighting by introducing a parameter ϕ_i for each X_i and rewriting the performance measure for X_i as follows:

$$\begin{aligned} &f(Z, c_Z; X_i; \frac{\beta_i}{\sigma_i}; c_{X_i}; \phi_i) = \\ &2\phi_i \left(c_Z^T (Z^T Z)^{-1} c_Z \left[1 + tr\left(P_Z \frac{X_i \beta_i \beta_i^T X_i^T}{(n-p)\sigma_i^2} \right) \right] \right) \\ &+ (2 - 2\phi_i) \left(\frac{1}{\sigma_i^2} c_Z^T (Z^T Z)^{-1} Z^T X_i \beta_i \beta_i^T X_i^T Z (Z^T Z)^{-1} c_Z \right) \\ &+ (2 - 2\phi_i) \left(-\frac{2}{\sigma_i^2} c_Z^T (Z^T Z)^{-1} Z^T X_i \beta_i \beta_i^T c_{X_i} \right) \\ &+ (2 - 2\phi_i) \left((c_{X_i}^T \beta_i / \sigma_i)^2 \right) \end{aligned} \tag{25}$$

The parameter $\phi_i \in (0, 1)$ controls the relative importance of the bias and variance terms. When $\phi_i = 0.5$, both terms are equally weighted as in 13. Higher values of ϕ_i give higher weight to the variance term and lower values of ϕ_i give higher weight to the bias term. The composite objective function for local bias-variance weighting is defined as before:

$$G_\phi(Z, c_Z) = \sum_{i=1}^{m} w_i f(Z, c_Z; X_i; \frac{\beta_i}{\sigma_i}; c_{X_i}; \phi_i) \tag{26}$$

Define diagonal matrices Φ_V and Φ_B as follows:

$$\Phi_V = \begin{pmatrix} 2\phi_1 & 0 & \dots \\ \vdots & \ddots & \vdots \\ 0 & \dots & 2\phi_m \end{pmatrix} \tag{27}$$

and

$$\Phi_B = \begin{pmatrix} 2 - 2\phi_1 & 0 & \dots \\ \vdots & \ddots & \vdots \\ 0 & \dots & 2 - 2\phi_m \end{pmatrix} \tag{28}$$

With these definitions G_ϕ can be written as:

$$\begin{aligned} G_\phi(Z, c_Z) &= c_Z^T (Z^T Z)^{-1} c_Z \left[\sum_{i=1}^{m} 2\phi_i w_i + tr\left(P_Z H \Phi_V \Sigma H^T \right) \right] \\ &+ c_Z^T (Z^T Z)^{-1} Z^T H \Phi_B H^T Z (Z^T Z)^{-1} c_Z \\ &- 2c_Z^T (Z^T Z)^{-1} Z^T H \Phi_B \ell \\ &+ \sum_{i=1}^{m} w_i (2 - 2\phi_i)(c_{X_i}^T \beta_i / \sigma_i)^2 \end{aligned} \tag{29}$$

This approach can be easily extended to the simultaneous optimization of multiple contrasts using the function:

$$G_\phi(Z, c_{Z_1}, \dots, c_{Z_q}) = \sum_{s=1}^{q} G_\phi(Z, c_{Z_s}) \tag{30}$$

E. Bias definitions for visualization

In this section, we define certain quantities designed to aid interpretation and visualization of properties of the computed optimal design matrices. These are by no means the only possible definitions.

Normalized contrast bias: How does $c_Z^T \hat{\gamma}$ compare with $c_X^T \beta$? To answer this question we define the normalized contrast bias C_b (assuming $\frac{c_X^T \beta}{\sigma} \neq 0$) (using 6) as follows:

$$\begin{aligned} C_b = \frac{c_Z^T E(\hat{\gamma}) - c_X^T \beta}{c_X^T \beta} &= \frac{(c_Z^T E(\hat{\gamma}) - c_X^T \beta)/\sigma}{(c_X^T \beta)/\sigma} \\ &= \frac{c_Z^T (Z^T Z)^{-1} Z^T X (\beta/\sigma)}{c_X^T (\beta/\sigma)} - 1 \end{aligned} \tag{31}$$

The motivation for this definition is that it is easy to think of bias as a fractional change in the PE of interest from the true value. With this definition, phrases like "1% bias" make intuitive sense.

When $\frac{c_X^T \beta}{\sigma} = 0$, normalization by $\frac{c_X^T \beta}{\sigma}$ results in division by 0 and thus C_b from 31 becomes illdefined. Therefore, in this case we skip the normalization by $c_X^T \beta / \sigma$ and define C_b as $(c_Z^T E(\hat{\gamma}) - c_X^T \beta)/\sigma$, i.e., $C_b = c_Z^T (Z^T Z)^{-1} Z^T X (\beta/\sigma)$. Small values of C_b mean either low % bias in the presence of signal ($c_X^T \beta / \sigma \neq 0$) or near 0 mean estimate $E(c_Z^T \hat{\gamma} / \sigma)$ in the absence of signal ($c_X^T \beta / \sigma = 0$).

Model variance bias: How does $\hat{\sigma}_1^2$ compare with σ^2? To answer this question, we define the normalized model variance bias V_b as follows:

$$V_b = E\left(\frac{\hat{\sigma}_1^2}{\sigma^2} \right) - 1 = \frac{\beta^T X^T P_Z X \beta}{(n-p)\sigma^2} \tag{32}$$

Small values of V_b mean that $\hat{\sigma}_1^2$ is a nearly unbiased estimator of σ^2 which is a desirable property.

Contrast variance change: How does the estimated variance of $c_Z^T \hat{\gamma}$ compare with the variance of the estimate $c_X^T \hat{\beta}$ (assuming the true model X is known)? To answer this question, we define the normalized contrast variance change with respect to the Gauss-Markov estimate as follows:

$$CV_\Delta = \frac{E(\hat{\sigma}_1^2/\sigma^2) c_Z^T (Z^T Z)^{-1} c_Z}{c_X^T (X^T X)^{-1} c_X} - 1 \tag{33}$$

If the true model X is known then the estimator $c_X^T\hat{\beta}$ of $c_X^T\beta$ will have the minimum possible variance for an unbiased linear estimator of β as per the Gauss-Markov theorem. We prefer small values of CV_Δ since they signify that the variance of $c_Z^T\hat{\gamma}$ is close to the optimal value.

Test statistic: It can be shown (see Proposition 9.2) that $\hat{\gamma}$ and $\hat{\sigma}_1^2$ are independent. Suppose we are interested in testing $\mathbf{H_0} : c_X^T\beta = 0$. Consider the test statistic:

$$T(\hat{\gamma},\hat{\sigma_1};Z;c_Z) = \frac{c_Z^T(\hat{\gamma}/\hat{\sigma_1})}{\sqrt{c_Z^T\,(Z^TZ)^{-1}\,c_Z}}$$
$$\sim \frac{\mathbf{N}\left(\frac{c_Z^T(Z^TZ)^{-1}Z^TX(\beta/\sigma)}{\sqrt{c_Z^T(Z^TZ)^{-1}c_Z}},1\right)}{\sqrt{\frac{\chi^2(n-p,\Delta)}{n-p}}} = \frac{\mathbf{N}\left(\frac{E(c_Z^T\hat{\gamma})/\sigma}{\sqrt{c_Z^T(Z^TZ)^{-1}c_Z}},1\right)}{\sqrt{\frac{\chi^2(n-p,\Delta)}{n-p}}} \tag{34}$$

Here the rightmost equality follows from 6. T will have small values under H_0 and large values when there is significant deviation from H_0. Receiver operating characteristic (ROC) curves can be obtained using this test statistic as described in section VI-B. An interesting case deserves special mention here. As described in Proposition 9.3 and 9.4 this case occurs when one column in X models the effect of interest and the other columns are nuisance columns i.e., $X = [x_1, X_2]$ where X_2 is the nuisance part of X. Suppose $Z = [z_1, X_2, Z_3]$ where z_1 and Z_3 are optimized but X_2 is fixed i.e., Z contains the nuisance part X_2 of X. Then as shown in 9.4 we have $E(c_Z^T\hat{\gamma}) = 0$ and $\Delta = 0$. Let us denote this special case by S^*. Consequently, from 34 the distribution of T under H_0 for this special case S^* will be

$$T(\hat{\gamma},\hat{\sigma_1};Z;c_Z|H_0,S^*) = \frac{\mathbf{N}\,(0,1)}{\sqrt{\frac{\chi^2(n-p,0)}{n-p}}} = t_{n-p} \tag{35}$$

where t_{n-p} represents a central Student-t distribution with $(n-p)$ degrees of freedom.

We also derive below a lower bound on the expected value of T statistic defined in 34 irrespective of whether H_0 is true or not. By the independence of $\hat{\gamma}$ and $\hat{\sigma}_1^2$ (see Proposition 9.2):

$$E\left(\frac{\hat{\gamma}}{\hat{\sigma}_1}\right) = E\left(\hat{\gamma}\right)E\left(\frac{1}{\sqrt{\hat{\sigma}_1^2}}\right) \tag{36}$$

Now the function $f(x) = \frac{1}{\sqrt{x}}$ is convex for $x > 0$ and thus by Jensen's inequality for any random variable V, $E\left(f\left(V\right)\right) \geq f\left(E\left(V\right)\right)$. Thus it follows that

$$E\left(\frac{1}{\sqrt{\hat{\sigma}_1^2}}\right) \geq \left(\frac{1}{\sqrt{E(\hat{\sigma}_1^2)}}\right) \tag{37}$$

From 36 and 37 we have:

$$E\left(\frac{\hat{\gamma}}{\hat{\sigma}_1}\right) = E\left(\hat{\gamma}\right)E\left(\frac{1}{\sqrt{\hat{\sigma}_1^2}}\right) \geq E\left(\hat{\gamma}\right)\left(\frac{1}{\sqrt{E(\hat{\sigma}_1^2)}}\right) \tag{38}$$

From 34 and 38 it follows that:

$$\begin{aligned} E(T) &= E\left(\frac{c_Z^T(\hat{\gamma}/\hat{\sigma_1})}{\sqrt{c_Z^T\,(Z^TZ)^{-1}\,c_Z}}\right) \\ &\geq \frac{c_Z^T(Z^TZ)^{-1}Z^TX(\beta/\sigma)}{\sqrt{c_Z^T\,(Z^TZ)^{-1}\,c_Z}\sqrt{1+\frac{\beta^TX^TP_ZX\beta}{(n-p)\sigma^2}}} \end{aligned} \tag{39}$$

III. Algorithm

In this section we address the following questions:

1) How to implement an algorithm for calculating optimal Z and c_Z? We discuss this in sections III-A-III-B.
2) How can we estimate the size of Z? How should we choose the bias-variance weights ϕ_i? This is discussed in sections III-C-III-D.

A. Implementation

In this section we describe simplified optimization strategy that seems to work for the nature of the problem under consideration. Basically it involves simple gradient descent steps with adaptive step sizes. This practical algorithm is summarized in Algorithm 1. In appendix IX-E, we validate this algorithm by comparing optimal solutions from the more sophisticated solver with the ones produced using Algorithm 1.

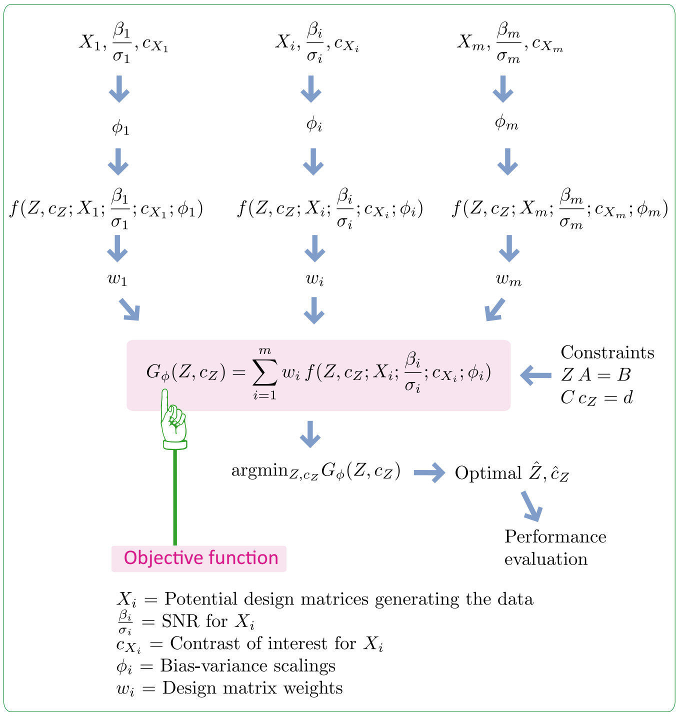

$X_1, \frac{\beta_1}{\sigma_1}, c_{X_1}$
$X_i, \frac{\beta_i}{\sigma_i}, c_{X_i}$
$X_m, \frac{\beta_m}{\sigma_m}, c_{X_m}$
ϕ_1
ϕ_i
ϕ_m
$f(Z,c_Z;X_1;\frac{\beta_1}{\sigma_1};c_{X_1};\phi_1)$
$f(Z,c_Z;X_i;\frac{\beta_i}{\sigma_i};c_{X_i};\phi_i)$
$f(Z,c_Z;X_m;\frac{\beta_m}{\sigma_m};c_{X_m};\phi_m)$
w_1
w_i
w_m
$G_\phi(Z,c_Z) = \sum_{i=1}^{m} w_i\, f(Z,c_Z;X_i;\frac{\beta_i}{\sigma_i};c_{X_i};\phi_i)$
Constraints
$Z\,A = B$
$C\,c_Z = d$
$\text{argmin}_{Z,c_Z} G_\phi(Z,c_Z)$
Optimal $\hat{Z}, \hat{c}_Z$
Performance evaluation
Objective function
X_i = Potential design matrices generating the data
$\frac{\beta_i}{\sigma_i}$ = SNR for X_i
c_{X_i} = Contrast of interest for X_i
ϕ_i = Bias-variance scalings
w_i = Design matrix weights

Fig. 1. Flowchart for computing optimal design matrices. Design matrices X_i, contrasts of interest c_{X_i} and signal to noise ratios $\frac{\beta_i}{\sigma_i}$ are locally weighted using bias-variance weighting ϕ_i to compute performance measures $f(Z, c_Z; X_i; \frac{\beta_i}{\sigma_i}; c_{X_i}, \phi_i)$. These performance measures are combined using weights w_i to form the objective function $G_\phi(Z, c_Z)$. Function $G_\phi(Z, c_Z)$ is then optimized with respect to Z and c_Z subject to user defined constraints to yield the optimal design matrices $\hat{Z}$ and $\hat{c_Z}$. A post-processing step then generates performance curves for user inspection.

Algorithm 1: Algorithm for optimizing DM

```
Require: Problem variables $H$, $\Sigma$, $\ell$, $A$, $C$
Require: Algorithmic variables $\alpha_0 \in (0, 10^{-3})$, $\theta \in (1, 5]$
    and $\eta_1, \eta_2 \in (0, 10^{-6})$
Require: The size of optimal DM, $p$ and Initial point $Z_0$, $c_{Z_0}$
    satisfying $Z_0 A = B$ and $Cc_{Z_0} = d$
Ensure: Outputs are the optimal DM, $\hat{Z}$, the optimal contrast,
    $\hat{c_Z}$ and the optimal objective function, $\hat{F}$
Compute orthogonal projectors $P_A = I_p - A(A^TA)^{-1}A^T$
    and $P_{C^T} = I_p - C^T(CC^T)^{-1}C$
$found = 0$, $j = 0$
while $found = 0$ do
    Let $S_j = \frac{\partial G}{\partial Z}(Z_j, c_{Z_j})$ and $T_j = \frac{\partial G}{\partial c_Z}(Z_j, c_{Z_j})$
    $success = 0$
    while $success = 0$ do
        $Z_{j+1} = Z_j - \alpha_j S_j P_A$
        $c_{Z_{j+1}} = c_{Z_j} - \alpha_j P_{C^T} T_j$
        $F_j = G(Z_j, c_{Z_j})$ and $F_{j+1} = G(Z_{j+1}, c_{Z_{j+1}})$
        if $F_{j+1} < F_j$ then
            $\alpha_{j+1} = \theta\alpha_j$
            $success = 1$
        else
            $\alpha_j = \alpha_j/\theta$
        end if
    end while
    if $||F_{j+1} - F_j|| \leq \eta_1$ or $\alpha_{j+1} \leq \eta_2$ then
        $found = 1$
    else
        $j = j + 1$
    end if
end while
return $\hat{Z} = Z_{j+1}$, $\hat{c_Z} = c_{Z_{j+1}}$ and $\hat{F} = F_{j+1}$
```

Fig. 2. Algorithm for optimizing DM

The gradients of G_ϕ are given by: (see appendix IX-C for detailed derivation)

$$\begin{aligned}
\frac{\partial G_\phi}{\partial Z} &= -Z(Z^TZ)^{-1}(2c_Zc_Z^T)(Z^TZ)^{-1}\left[\sum_{i=1}^{m} 2\phi_i w_i\right] \\
&- Z(Z^TZ)^{-1}(2c_Zc_Z^T)(Z^TZ)^{-1}\left[tr\left(P_Z H\Phi_V \Sigma H^T\right)\right] \\
&- 2(c_Z^T(Z^TZ)^{-1}c_Z)P_Z H\Phi_V\Sigma H^T Z(Z^TZ)^{-1} \\
&- 2Z(Z^TZ)^{-1}c_Zc_Z^T(Z^TZ)^{-1}Z^T H\Phi_B H^T Z(Z^TZ)^{-1} \\
&- 2Z(Z^TZ)^{-1}(Z^TH\Phi_BH^TZ)(Z^TZ)^{-1}c_Zc_Z^T(Z^TZ)^{-1} \\
&+ 2H\Phi_BH^TZ(Z^TZ)^{-1}c_Zc_Z^T(Z^TZ)^{-1} \\
&+ 2Z(Z^TZ)^{-1}Z^TH\Phi_B\ell c_Z^T(Z^TZ)^{-1} \\
&+ 2Z(Z^TZ)^{-1}c_Z\ell^T\Phi_B^TH^TZ(Z^TZ)^{-1} \\
&- 2H\Phi_B\ell c_Z^T(Z^TZ)^{-1}
\end{aligned} \tag{40}$$

$$\begin{aligned}
\frac{\partial G_\phi}{\partial c_Z} &= 2(Z^TZ)^{-1}c_Z\left[\sum_{i=1}^{m} 2\phi_i w_i + tr\left(P_Z H\Phi_V\Sigma H^T\right)\right] \\
&+ 2(Z^TZ)^{-1}Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Z \\
&- 2(Z^TZ)^{-1}Z^TH\Phi_B\ell
\end{aligned}$$

When optimizing over multiple contrasts as per 30, the gradients are given by:

$$\frac{\partial G_\phi}{\partial Z} = \sum_{s=1}^{q} \frac{\partial G_\phi(Z, c_{Z_s})}{\partial Z} \text{ and } \frac{\partial G_\phi}{\partial c_{Z_r}} = \frac{\partial G_\phi(Z, c_{Z_r})}{\partial c_{Z_r}} \tag{41}$$

B. Note on initialization

It is well known that when finding a local solution to an optimization problem as we do here, the choice of an initial point could have an impact on the estimated local solution. We acknowledge that there could be many interesting initialization strategies. Here we propose one such strategy for initialization of Z and c_Z. We recommend a heuristic strategy for initialization of the primary column in Z. First we try to find a vector v that is closest to primary columns in X_i in the following sense:

$$\text{argmin}_{\,v} \sum_{i=1}^{m} ||\left(c_{X_i}^T\frac{\beta_i}{\sigma_i}\right)(X_i c_{X_i}) - \left(c_{X_i}^T\frac{\beta_i}{\sigma_i}\right)v)||_2^2 \tag{42}$$

The solution to 42 is given by:

$$\hat{v} = \frac{\sum_{i=1}^{m}\left(c_{X_i}^T\frac{\beta_i}{\sigma_i}\right)(X_i c_{X_i})}{\sum_{i=1}^{m}\left(c_{X_i}^T\frac{\beta_i}{\sigma_i}\right)} \tag{43}$$

Next we form the $n \times m$ matrix $M = [M_1, \ldots, M_i, \ldots M_m]$ of residuals with ith column:

$$M_i = \left[\left(c_{X_i}^T\frac{\beta_i}{\sigma_i}\right)(X_i c_{X_i}) - \left(c_{X_i}^T\frac{\beta_i}{\sigma_i}\right)\hat{v}\right] \tag{44}$$

Next we take the singular value decomposition of M to get:

$$M = U_M\Sigma_M V_M^T \tag{45}$$

where $U_M \in \mathbf{R}^{n\times m}$ is a matrix of left singular vectors, $V_M \in \mathbf{R}^{m\times m}$ is a matrix of right singular vectors and Σ_M holds the singular values of M.

For optimizing a p column matrix Z we choose (in matlab notation) the following initialization:

$$Z_0(:,1) = \hat{v} \tag{46}$$

$$Z_0(:,2:p) = U_M(:,1:(p-1)) \tag{47}$$

$$c_{Z_0} = [1; 0; \ldots; 0] \text{ (p rows)} \tag{48}$$

Thus the vector $\hat{v}$ is used to initialize the primary column of Z_0 and the columns of U_M are used to initialize the non-primary columns of Z_0. The next step is to modify Z_0 and c_{Z_0} so that they satisfy the constraints $Cc_{Z_0} = d$ and $Z_0A = B$.

$$Z_0 = Z_0 + (B - Z_0A)(A^TA)^{-1}A^T \tag{49}$$

$$c_{Z_0} = c_{Z_0} + C^T(CC^T)^{-1}(d - Cc_{Z_0}) \tag{50}$$

```
**Algorithm 2:** Choosing optimal number of columns in $Z$

Require: Problem variables $H$, $\Sigma$, $\ell$, $A$, $C$
Require: Algorithmic variables $\alpha_0 \in (0, 10^{-3})$, $\theta \in (1, 5]$ and $\eta_1, \eta_2 \in (0, 10^{-6})$
Require: Initial choice $p = p_0$ and $p_{max}$ the maximum value of $p$
Require: User chosen accuracy cutoff $R_c \in (0.5, 1)$ with a default value of $R_c = 0.95$ (95% cutoff)
Ensure: Outputs are the optimal number of columns in the DM $p_{opt}$
Set $j = 1$
for $p = p_0$ to $p_{max}$ do
    Require: $Z_0$, $c_{Z_0}$ satisfying $Z_0 A = B$ and $C c_{Z_0} = d$ such that $Z_0$ and $c_Z$ have $p$ columns
    Run Algorithm 1 to estimate $\hat{Z}$, $\hat{c_Z}$ and $\hat{F}$ for the current value of $p$
    Set $p_{est}(j) = p$
    Set $F(j) = \hat{F}$, $Z(j) = \hat{Z}$ and $c_Z(j) = \hat{c_Z}$
    $j = j + 1$
end for
Compute $F_{max} = max_j F(j)$ and $F_{min} = min_j F(j)$
Set $j = 1$
for $p = p_0$ to $p_{max}$ do
    $R(j) = (F_{max} - F(j))/(F_{max} - F_{min})$
    $j = j + 1$
end for
Calculate the minimum $j$ meeting the cutoff, $\hat{j} = min\{j : R(j) \geq R_c\}$
return $p_{opt} = p_{est}(\hat{j})$
```

Fig. 3. Choosing optimal number of columns in Z

This is not the only way to initialize Z, there can be many other strategies. In fact we have not used this initialization strategy in many of the examples presented here precisely to illustrate this point.

C. Estimating the size of Z

By accounting for expected variations in the shape and size of the response and then optimizing for a design matrix Z via the solution of an inverse problem that explicitly controls for bias and variance we automatically avoid overfitting in this framework. The framework also allows for inclusion of ”null” data that is not simply Gaussian noise but is some structured signal such as a drift (see Example 3) to explicitly instruct the optimization process to ”equate” it to ”no signal” during optimization. Why then is it important to choose the size of Z? One reason is to maximize the degrees of freedom available for subsequent first level or higher level statistical tests. We propose the following strategy for choosing the ”optimal” number of columns in Z.

```
**Algorithm 3:** Choosing optimal number of columns in $Z$ - sensitivity to initialization

Require: Problem variables $H$, $\Sigma$, $\ell$, $A$, $C$
Require: Algorithmic variables $\alpha_0 \in (0, 10^{-3})$, $\theta \in (1, 5]$ and $\eta_1, \eta_2 \in (0, 10^{-6})$
Require: Initial choice $p = p_0$ and $p_{max}$ the maximum value of $p$
Require: User chosen accuracy cutoff $R_c \in (0.5, 1)$ with a default value of $R_c = 0.95$ (95% cutoff)
Require: User chosen number of trials $n_{iter}$
Ensure: Outputs are the optimal number of columns in the DM $p_{opt}(n_{iter})$ over $n_{iter}$ runs
for $j = 1$ to $n_{iter}$ do
    Run Algorithm 2 to get $p_{opt}$ for this run
    Set $p_{est}(j) = p_{opt}$
end for
$p_{opt}(n_{iter}) = \text{Median}_j p_{est}(j)$
return $p_{opt}(n_{iter})$
```

Fig. 4. Choosing optimal number of columns in Z - Sensitivity to Initialization

The basic idea in Algorithm 2 is to run Algorithm 1 for a range of values of p and choose a value of p that achieves a user chosen reduction in the objective function value relative to the maximum possible reduction over all values of p. Please note that the strategy for choosing the number of columns proposed here is by no means the only one. For example, it could also involve reduction in the model variance V_b below a specified user value. If the R vs p curve has a local maximum as opposed to a monotonic increase then the location of this maximum also is a reasonable choice for the optimal size of Z. In principle one can correct for variability due to initialization using Algorithm 3 that essentially runs Algorithm 2 for a number of initializations.

We ran Algorithm 2 for Validation Test A (see Appendix IX-E1) and the results are shown in Figures 5-6. It was found that $p_{opt} = 3$ using a cutoff of $R_c = 0.95$ for this case study.

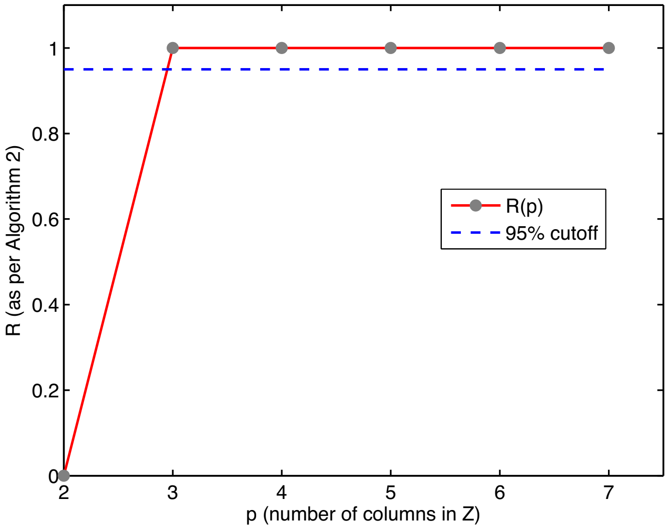


Fig. 6. Data from Validation Test A (see Appendix IX-E1) showing R versus p curve and the cutoff at 0.95 indicating that $p_{opt} = 3$

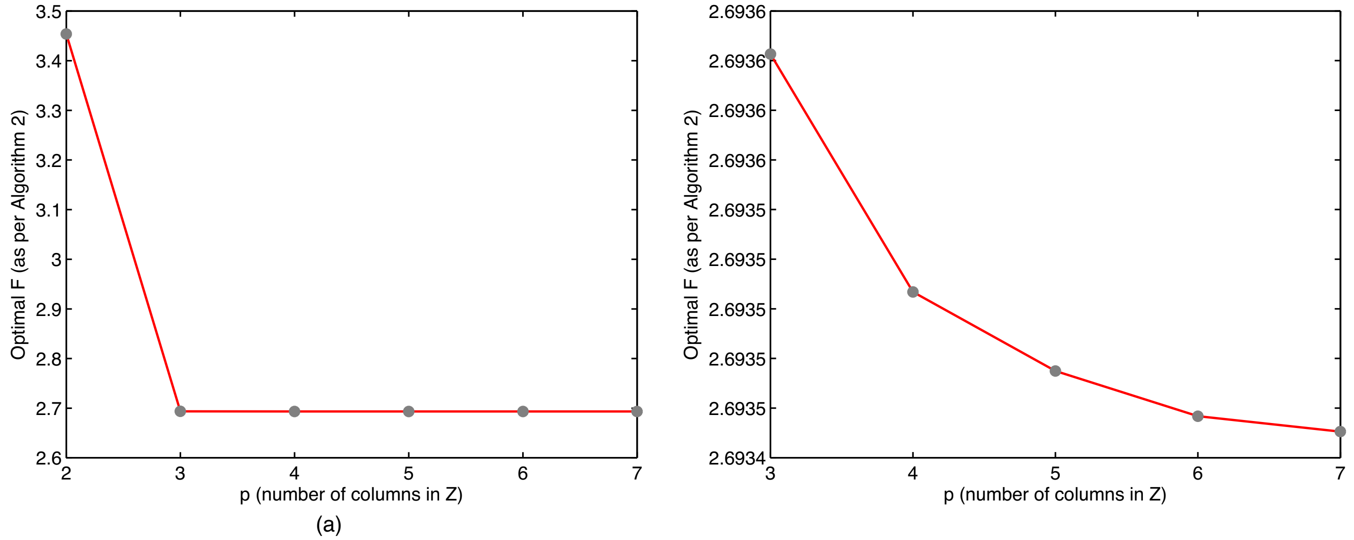


Fig. 5. Illustration of the procedure to estimate the size of Z using data from Validation Test A (see Appendix IX-E1) (a) Number of columns p versus the optimal objective $\hat{F}$ (b) Same as (a) but showing the slow decline of the tail in (a)

D. Choice of ϕ_i

So far we have not discussed sensible choices of ϕ_i for design matrix X_i. Is a constant ϕ_i for all i the best choice? In this section we wish to motivate some heuristic rules for selecting ϕ_i based on user defined objectives. Equation 25 can be re-written using definitions of C_b and CV_Δ for X_i from 31 and 33 as follows:

$$\begin{aligned} f(Z, c_Z; X_i; \frac{\beta_i}{\sigma_i}; c_{X_i}; \phi_i) = \\ \left\{(2\phi_i) c_{X_i}^T (X_i^T X_i)^{-1} c_{X_i}\right\} (CV_{\Delta i} + 1) \\ + \left\{ (2 - 2\phi_i) \left(\frac{c_{X_i}^T \beta_i}{\sigma_i} \right)^2 \right\} C_{bi}^2 \end{aligned} \quad (51)$$

The contribution of fractional contrast bias C_{bi}^2 to the objective function is controlled by its multiplier $\left\{ (2 - 2\phi_i) \left(\frac{c_{X_i}^T \beta_i}{\sigma_i} \right)^2 \right\}$. Similarly, the contribution of fractional variance change w.r.t Gauss-Markov estimate term $(CV_{\Delta i} + 1)$ is controlled by its multiplier $\left\{(2\phi_i) c_{X_i}^T (X_i^T X_i)^{-1} c_{X_i}\right\}$.

1) Choice A: Suppose the user wants to put k times more emphasis on the C_{bi}^2 term compared to $(CV_{\Delta i} + 1)$ term. Then a sensible choice would be:

$$\left\{ (2 - 2\phi_i) \left(\frac{c_{X_i}^T \beta_i}{\sigma_i} \right)^2 \right\} = k \left\{(2\phi_i) c_{X_i}^T (X_i^T X_i)^{-1} c_{X_i}\right\} \quad (52)$$

Solving for ϕ_i and denoting the calculated value by $\phi_i(k)$ to indicate k times more emphasis on C_{bi}^2 term compared to $(CV_{\Delta i} + 1)$ term, we get:

$$\phi_i(k) = \frac{\left(\frac{c_{X_i}^T \beta_i}{\sigma_i} \right)^2}{\left(\frac{c_{X_i}^T \beta_i}{\sigma_i} \right)^2 + k\, c_{X_i}^T (X_i^T X_i)^{-1} c_{X_i}} \quad (53)$$

It is clear that as $k \to \infty$, $\phi_i(k) \to 0$ and as $k \to 0$, $\phi_i(k) \to 1$. This behavior is as expected but interestingly it is non-linear. In particular, choosing $k = 1$ i.e., equal emphasis on C_{bi}^2 and $(CV_{\Delta i} + 1)$ terms does not necessarily imply $\phi_i(1) = 0.5$. Choosing sensible values of k boils down to sifting through the optimal solutions and picking ones that make practical sense. In general one wants $C_{bi}^2 \sim 10^{-3}$ and $CV_\Delta \sim 0$. Thus it makes sense to choose $k = 10^3$ to make both terms in the optimization of a similar magnitude while satisfying reasonable practical objectives.

2) Choice B: Suppose the user wants to put k times more emphasis on the $|C_{bi}|$ term compared to $(CV_{\Delta i} + 1)$ term. In this case a sensible choice would be:

$$\sqrt{(2 - 2\phi_i)} \left| \left(\frac{c_{X_i}^T \beta_i}{\sigma_i} \right) \right| = k \left\{(2\phi_i) c_{X_i}^T (X_i^T X_i)^{-1} c_{X_i}\right\} \quad (54)$$

Squaring both sides and solving the resulting quadratic equation and discarding the negative root results in:

$$\phi_i(k) = \frac{-1 + \sqrt{1 + 4\, a(k)}}{2\, a(k)} \quad (55)$$

where

$$a(k) = \frac{2 \left\{ c_{X_i}^T (X_i^T X_i)^{-1} c_{X_i} \right\}^2 k^2}{\left(\frac{c_{X_i}^T \beta_i}{\sigma_i} \right)^2} \quad (56)$$

It can be shown that as $k \to \infty, a \to \infty, \phi_i(k) \to 0$ and as $k \to 0, a \to 0, \phi_i(k) \to 1$. Again, this behavior is to be expected. As before, choosing $k = 1$ i.e., equal emphasis on $|C_{bi}|$ and $(CV_{\Delta i} + 1)$ terms does not imply $\phi_i(1) = 0.5$. Practically speaking one wants $|C_{bi}| \sim 10^{-2}$ and $CV_\Delta \sim 0$ and hence a reasonable value of k for choice B would be $k = 10^2$ to make the two terms comparable in magnitude during the optimization process.

An example of the non-linear relationship between $\log k$ and ϕ is shown in Figure 7.

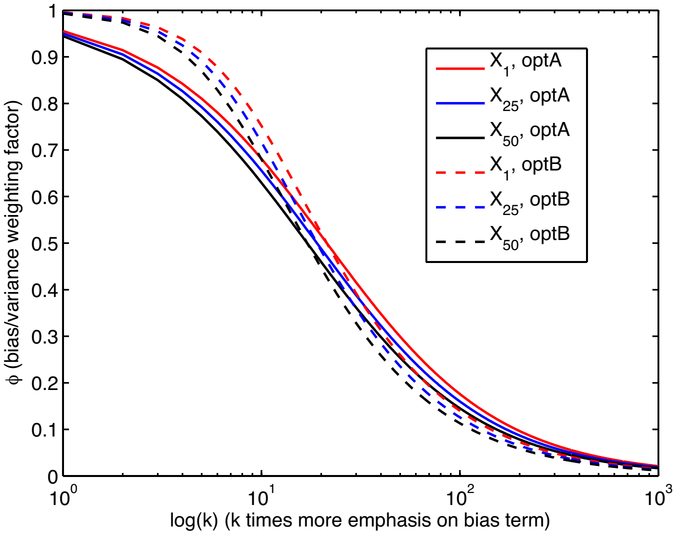


Fig. 7. Plot of $\log k$ vs ϕ for design matrices X_1, X_{25} and X_{50} from Example 1. This Figure shows that as $k \longrightarrow \infty$, $\phi \longrightarrow 0$ and as $k \longrightarrow 0$, $\phi \longrightarrow 1$ but the relationship is non-linear, i.e., choosing $\phi = 0.5$ does not give equal emphasis to bias and variance terms ($k \neq 1$).

IV. Case Studies

In this section we present performance curves for the estimated optimal DM's for six example cases. In particular, we look at the contrast bias C_b, the model variance bias V_b and contrast variance change w.r.t the Gauss-Markov estimate CV_Δ as key performance measures.

In the first 3 examples, X_i, $\frac{\beta_i}{\sigma_i}$, c_{X_i} were chosen as in IV-A. In the Example 4 additional X_i were added to those in Examples 1-3 for illustration purposes. Example 5 deals with a new set of X_i derived from the standard block design used in fMRI experiments. Example 6 illustrates the application of proposed technique to the problem of capturing variable shapes of the Haemodynamic response function (HRF) in fMRI.

A. Example 1

Motivated by a practical data-set that we later describe we consider for illustration purposes the case study when the profiles of interest are shifted relative to a base profile by variable units and our goal is to simultaneously capture all responses with a single design matrix. To test and validate the optimization framework, we used the basic design matrix X_0 from Figure 20. Fifty ($m = 50$) expected design matrices were proposed with

$$X_i(:,1) = X_0(:,1) \text{ shifted right by } i \text{ timepoints} \tag{57}$$

$$X_i(:,2) = X_0(:,2) \tag{58}$$

We chose $\frac{\beta_i}{\sigma_i} = [1, 0.5]^T$ and $c_{X_i} = [1, 0]^T, \forall i$. The weights were chosen as $w_i = 1, \forall i$ to reflect the equal likelihood of observing any X_i. In this example, we used the default weighting of bias and variance by choosing $\phi_i = 0.5$.

The rank of Z was chosen to be 3 and the matrix A was chosen as $A = [e_1, e_2]$ where $e_k \in R^3$ is a unit vector with 1 at position k and zeros elsewhere. Similarly for e_2. The matrix B was chosen as X_0 to fix the first two columns of Z to those of X_0. C was chosen as the identity matrix I_3 and d was set to $[1, 0, 0]^T$ to fix the contrast c_Z. The unconstrained column in Z was initialized randomly with elements drawn from a uniform distribution $U(0, 1)$. The results are shown in Figures 8 and 9.

B. Example 2

This example is the same as Example 1 except for two differences:

1) First we choose ϕ_i automatically using the strategy proposed in section III-D. We used Option A to initialize ϕ_i using $k = 10^3$.
2) Second, we initialize the columns of Z automatically using the strategy proposed in section III-B.

The results are shown in Figures 27 and 28 in appendix IX-D1.

C. Example 3

In this example, we attempt to investigate the effect of bias-variance weightings on the performance curves. We put a higher emphasis on reducing bias by choosing $\phi = 0.1$. Please note that the relationship between the value of ϕ_i and the importance of bias or variance terms is non-linear (see section III-D). The first two columns of Z were fixed as before and the contrast c_Z was fixed at [1;0;0]. We also chose $w_i = 1$ to give equal weights to all design matrices. The unconstrained column in Z was initialized randomly with elements drawn from a uniform distribution $U(0, 1)$. The results are shown in Figures 29 and 30 in appendix IX-D2.

D. Example 4

In this example, we attempt to investigate the effect of bias-variance weightings on the performance curves as well as the effect of optimizing the entire design matrix Z. We put a much higher emphasis on reducing bias by choosing $\phi = 0.01$. As before, note that the relationship between the value of ϕ_i and the importance of bias or variance terms is non-linear (see section III-D). The contrast c_Z was fixed at [1;0;0] but the design matrix Z was left unconstrained. We also chose $w_i = 1$ to give equal weights to all design matrices. The potential DM set from before was augmented with the following additional DM's

1) X_i, c_{X_i} same as before but $\frac{\beta_i}{\sigma_i} = [-1; 0.5]$
2) X_i, c_{X_i} same as before but $\frac{\beta_i}{\sigma_i} = [1; -0.5]$
3) X_i, c_{X_i} same as before but $\frac{\beta_i}{\sigma_i} = [-1; -0.5]$
4) X_0, c_{X_0} same as before but $\frac{\beta_0}{\sigma_0} = [0; 1]$

This was done to constrain the sample space to make Z unbiased for sign changes relative to drift (1,2, and 3) as well as unbiased for pure drift (4). The 3rd column of Z was initialized to the optimal solution found in Example 1. The results are shown in Figures 10 and 11.

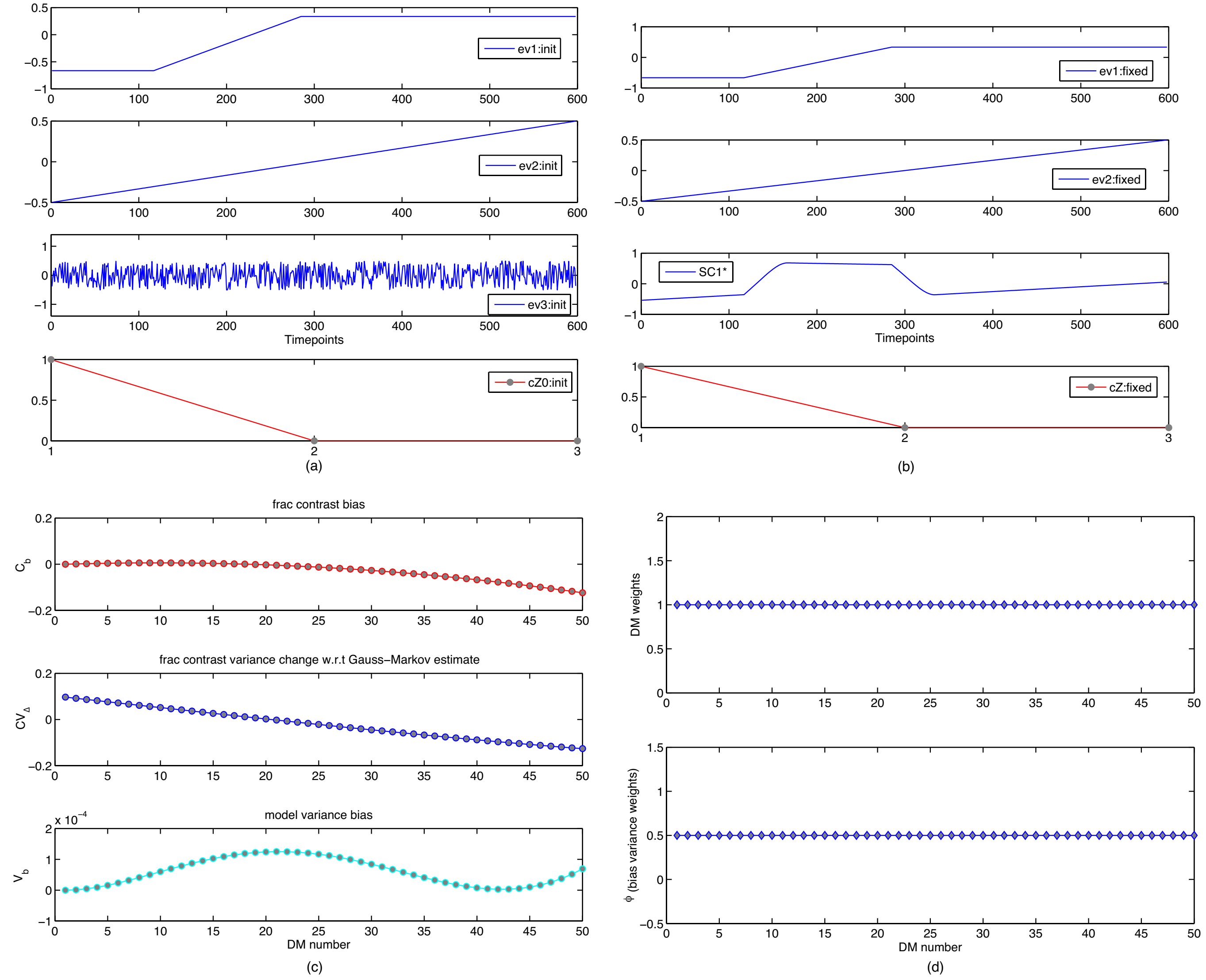


Fig. 8. Example 1: (a) Initial design matrix (DM) along with random initialization of the 3rd column. The first two columns were fixed at their initial values and the contrast was fixed at [1;0;0] (b) Estimated optimal DM. Notice how the 3rd column converges to a non-random profile (c) Performance curves showing the fractional contrast bias C_b, contrast variance change w.r.t Gauss-Markov estimate $CV_\triangle$ and model variance bias V_b (d) For all i, the DM weights $w_i = 1$ implying equal likelihood of observing any of the specified DM's and bias-variance scalings $\phi_i = 0.5$ for all i implying the default weighting of bias and variance.

E. Example 5

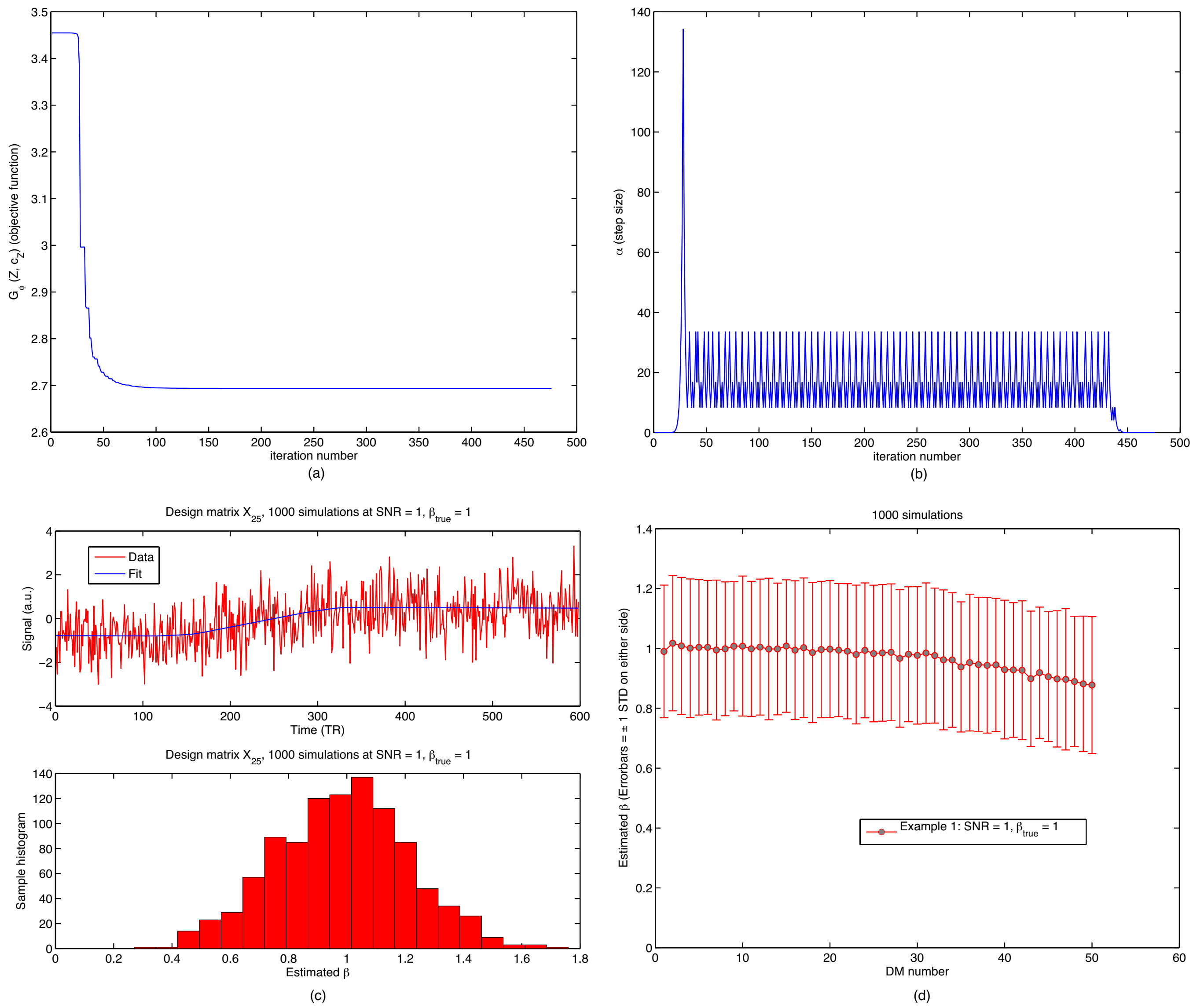
We use a standard block design EV to illustrate application of the proposed technique. The EV consists of blocks of 0's and 1's (see 12). The data might contain shifted or unshifted responses. The user estimates a maximum shift of 6 timepoints and would like to capture the variable data using an optimized design matrix. The user predominantly wants to control bias when the entire matrix Z is optimized for a fixed contrast $c_Z = [1;0;0;0]$. It is also desired to keep the main EV fixed as per the experimental paradigm and that the optimized contrast yield an unbiased estimate both for "positive" and "negative" activation. For this example, we chose $p = 4$ for the size of the optimal DM, the weights $w_i = 1$ and bias variance weights $\phi_i = 0.01$. The set of potential DM's in this case capturing both "positive" and "negative" responses is thus:

1) For $i = 1, \ldots, 6$: X_i = basic block design EV shifted to the right by i timepoints, c_{X_i} = [1] and $\frac{\beta_i}{\sigma_i}$ = [1]
1) For $i = 7, \ldots, 12$: $X_i = X_{i-6}$, c_{X_i} = [1] and $\frac{\beta_i}{\sigma_i}$ = [-1]

The results are shown in Figures 12 and 13.

F. Example 6

It is well known that there is no single haemodynamic response function (HRF) that captures the impulse response properties of all voxels in the brain. Here we illustrate the application of the technique developed in this paper to enable the simultaneous capture of a set of plausible HRF shapes. Plausible HRF shapes can be generated using any reasonable parameterization of HRF. Here we generate HRF shapes using the 5 parameter half-cosine parameterization as used in [9]. The details of this parameterization of HRF are given in appendix IX-G. 200 plausible HRF shapes were generated by sampling the 5 parameters from a uniform distribution as



(a)

(b)

(c)

(d)

Fig. 9. Example 1: (a) Figure showing the evolution of objective function values $G_\phi(Z, c_Z)$ over algorithm iterations. Notice how the function value stabilizes as convergence is reached (b) Figure showing the variation in the step size α over algorithm iterations. Step size controlling parameter θ in Algorithm 1 was set to $\theta = 2$. For each design matrix (DM) X_i entered into optimization, 1000 simulated data-sets were generated at SNR $\frac{\beta_i}{\sigma_i}$. A GLM analysis was run on each of these data-sets using the optimized DM. Figure (c) shows an example of simulated data for DM X_{25} at SNR $\frac{\beta_{25}}{\sigma_{25}}$ and the GLM fit using the optimal DM. It also shows the distribution of $c_Z^T\hat{\gamma}$ over 1000 simulations. Figure (d) is a summary errorbar plot showing $\hat{E}(c_Z^T\hat{\gamma})$ over 1000 simulations for data generated from each DM. The error bars represent unit standard deviation of $c_Z^T\hat{\gamma}$ (**not** standard deviation of $\hat{E}(c_Z^T\hat{\gamma})$) to quantify the variance in estimation via simulation.

follows:

$$\begin{aligned} h_1 &= U(1s, 3s) \\ h_2 &= U(3s, 7s) \\ h_3 &= U(3s, 7s) \\ h_4 &= U(3s, 9s) \\ f &= U(0, 0.5) \end{aligned} \qquad (59)$$

The units of $h_1, \ldots, h_4$ are in sec (s) and f is dimensionless. $U(a, b)$ denotes the uniform distribution on $[a, b]$. Data was generated by sampling at every 0.1 s.

The automatic initialization strategy described before was used for initialization. The weights w_i were set to 1 and ϕ_i were set to their default values of 0.5. Optimization results are shown in Figures 15 and 16. Once the set of ”optimal” HRF capturing functions are found, they can be entered into any GLM analysis as follows:

1) Convolve the experimental EV with each of the ”optimal” HRF capturing functions.
2) Next enter the resulting DM into a GLM analysis and use the contrast c_Z from the optimization above to capture the ”size” of underlying signal optimally.

V. Application to real fMRI data

In this section, we describe a case study in detail and show how the approach can be applied in practice.

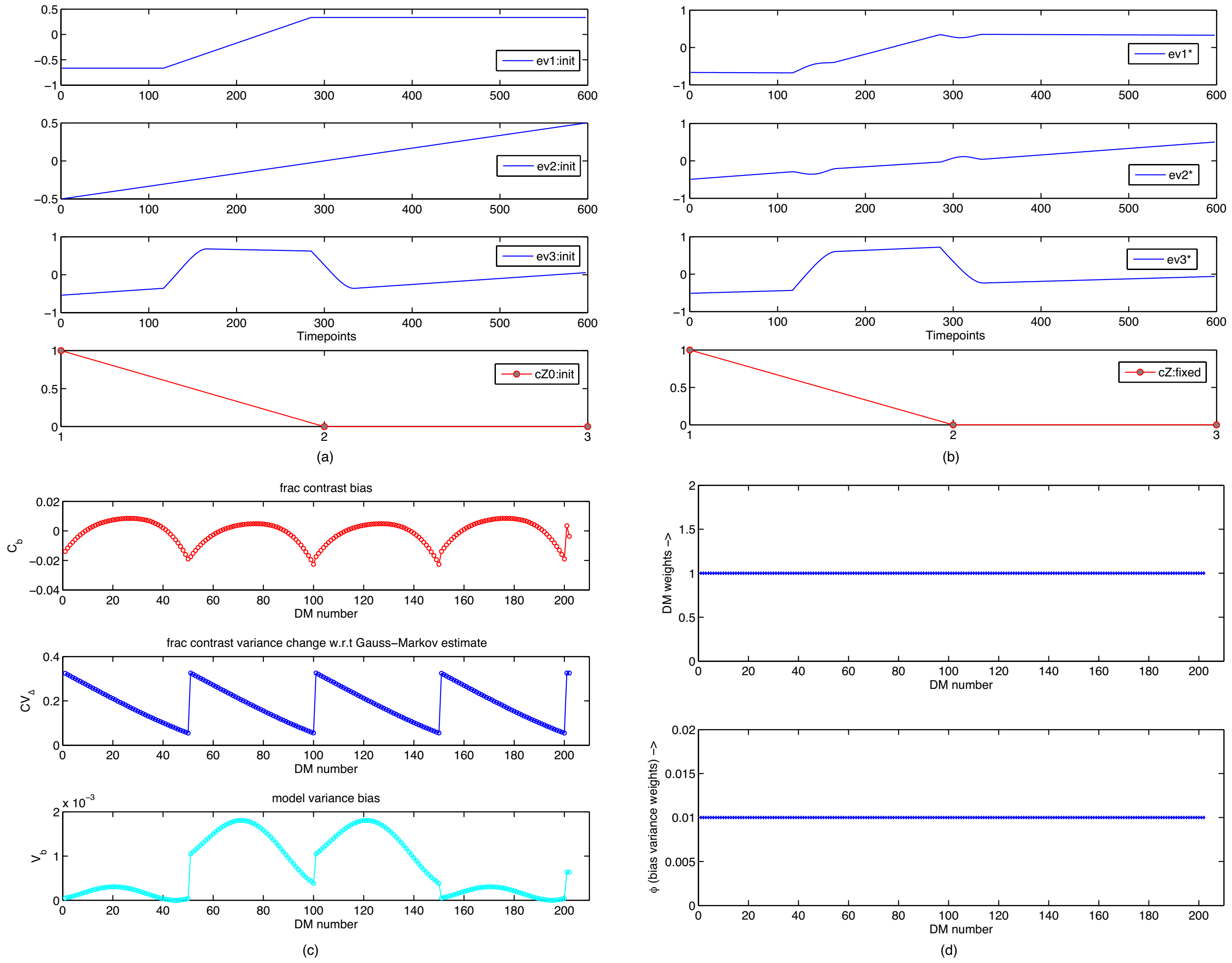


Fig. 10. Example 4: (a) Initial design matrix (DM). The 3rd column was initialized using the 3rd col of $\hat{Z}$ found in Example 1. The first two columns were left unconstrained in this case but the contrast was fixed at [1;0;0]. The set of potential DM's from Example 1 and 2 was augmented by adding more DM's at SNR's of $[1;-0.5], [-1;0.5], [-1;-0.5]$. A "null" DM was also added at SNR $[0;1]$ and $[0;-1]$ indicating that "pure drift" should be matched to size "0". (b) Estimated optimal DM. Notice how the unconstrained columns 1 and 2 converge to non-intuitive shapes (c) Performance curves showing the fractional contrast bias C_b, contrast variance change w.r.t Gauss-Markov estimate CV_Δ and model variance bias V_b (d) In this example $w_i = 1$ and $\phi_i = 0.01$ indicating a higher weight to the bias term during optimization.

A. Data acquisition

This case study describes a buprenorphine infusion scan at 0.2 mg/70kg dosage. Elimination half-lives for intravenous administration of buprenorphine (0.3 mg) range between 1.2 to 7.2 hours (mean = 2.2 hours), while the terminal half-life is 3 hours (Bullingham, et al. 1980). Onset of buprenorphine (intravenous, 0.3 mg to 0.6 mg dose) is within five minutes and analgesic effects lasts for 6 to 8 hours (Downing, et al. 1977).

In the 25 minute infusion scan, a 5 minute baseline was collected prior to the first infusion of buprenorphine or placebo. Four infusions, totaling 8 ml, were performed at minutes 5, 7, 9 and 11. Each 2 ml infusion was performed at a rate of 0.1 ml/sec and controlled by an automatic microinjector (Medrad Spectris, Colombus, OH).

All data were collected on a 3 Tesla Siemens Trio scanner with an 8-channel phased array head coil (Erlangen, Germany). Infusion data were collected using a gradient echo-echo planar pulse sequence (GE-EPI) at a 3.5 x 3.5 x 3.5 mm3 resolution. GE-EPI Parameters: Time of Repetition (TR) = 2500 msecs, Time of Echo (TE) = 30 msecs, Field of View (FOV) = 224x224, Flip Angle (FA) = 90, num of Slices = 41 axial slices and num of Volumes = 600. The acquisition time for the infusion scan was 25 mins and 5 secs. T1-weighted structural images were acquired using a 3-D magnetization-prepared rapid gradient echo (MPRAGE) sequence at a resolution of 1.33 x 1.0 x 1.0 mm3. MPRAGE Parameters: TR = 2100 msecs, TE = 2.74 msecs, Time of Inversion (TI) = 1100 msecs, FA = 12, num of Slices = 128 sagittal slices (Mugler and Brookeman 1990).

B. Data Analysis

1) Preprocessing: Single subject data analysis was performed using FMRIB Software Library (FSL)

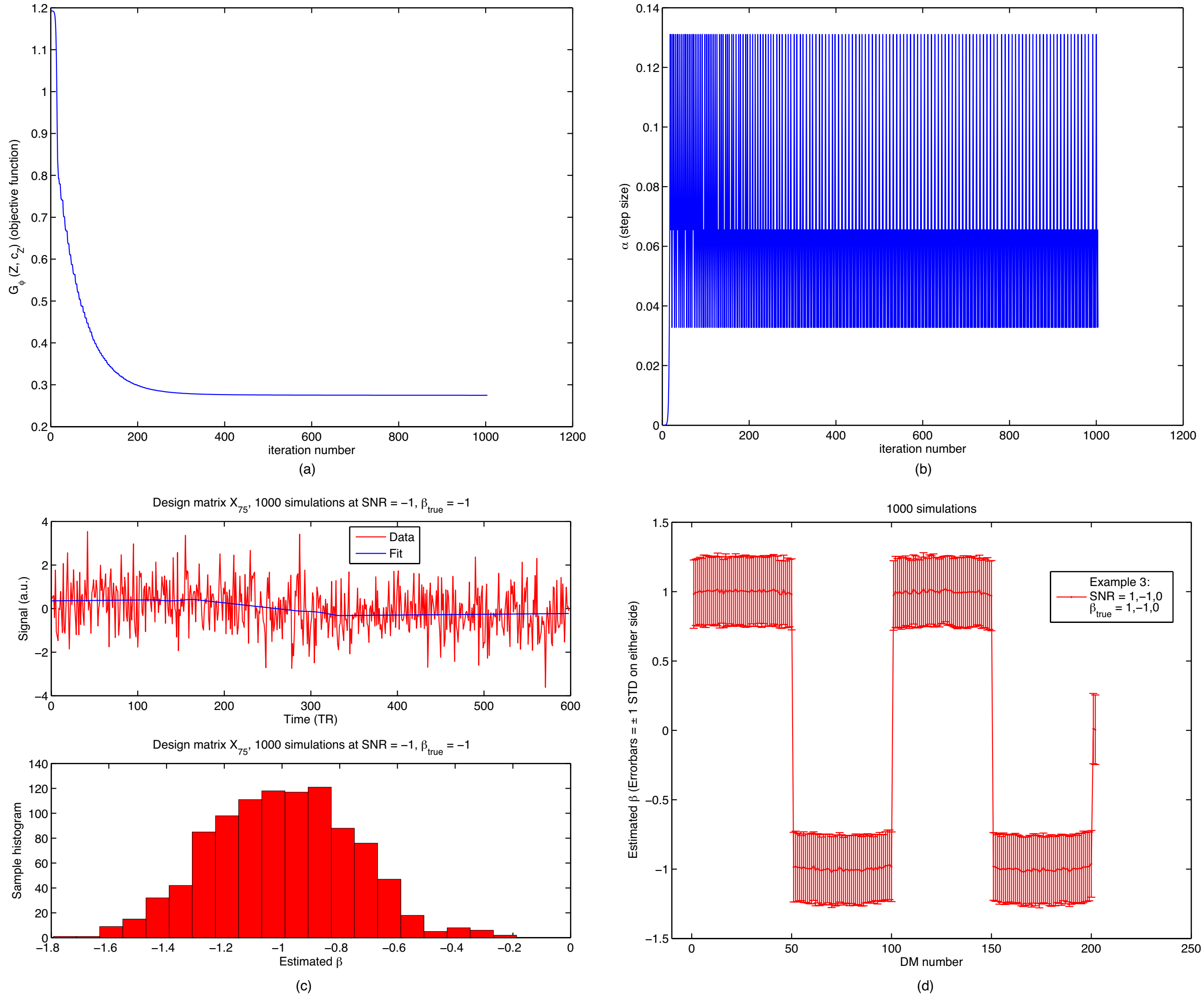


Fig. 11. Example 4: (a) Figure showing the evolution of objective function values $G_\phi(Z, c_Z)$ over algorithm iterations. Notice how the function value stabilizes as convergence is reached (b) Figure showing the variation in the step size α over algorithm iterations. Step size controlling parameter θ in Algorithm 1 was set to $\theta = 2$ (c) , (d) For each design matrix (DM) entered into optimization, 1000 simulated data-sets were generated at SNR $\frac{\beta_i}{\sigma_i}$. A GLM analysis was run on each of these data-sets using the optimized DM. Figure (c) shows an example of simulated data for DM X_{75} at SNR $\frac{\beta_{75}}{\sigma_{75}}$ and the GLM fit using the optimal DM. It also shows the distribution of $c_Z^T \hat{\gamma}$ over 1000 simulations. Figure (d) is a summary errorbar plot showing $\hat{E}(c_Z^T \hat{\gamma})$ over 1000 simulations for data generated from each DM. The error bars represent unit standard deviation of $c_Z^T \hat{\gamma}$ (**not** standard deviation of $\hat{E}(c_Z^T \hat{\gamma})$) to quantify the variance in estimation via simulation.

(http://www.fmrib.ax.ac.uk/fsl). The first two volumes were removed to account for MR signal instability in the initially acquired volumes. Raw fMRI data was preprocessed using the following steps 1) Skull stripping using a Brain Extraction Tool, 2) Motion correction using FMRIBs Linear Motion Correction tool (MCFLIRT), 3) Spatial smoothing with a 5 mm FWHM spatial filter.

2) Model Free analysis: Following this preprocessing, a projection pursuit analysis using the algorithm ADIS (Automated Decomposition Into Sources) was performed to extract spatial maps of minimum entropy to identify patterns of activity in the brain. ADIS (http://arxiv.org/abs/0902.4879, arXiv:0902.4879v1 [stat.CO]) is a probabilistic and constrained projection pursuit software that outperforms conventional ICA algorithms in several benchmark tests. Data was analyzed as follows:

1) Dimension reduction via probabilistic PCA. A lower bound for the latent dimension was determined via a bootstrap approach. A cross-validation analysis was used to estimate the true latent dimension.
2) The PCA-reduced data was decomposed into a set of minimum entropy spatial maps and their associated time-courses using ”negentropy” based projection pursuit. The ADIS optimization core was used to perform constrained optimization.The resulting z-maps were thresholded at $z > 3$ and $z < -3$.

The purpose of running ADIS was to demonstrate the existence of multiple infusion response profiles even for single

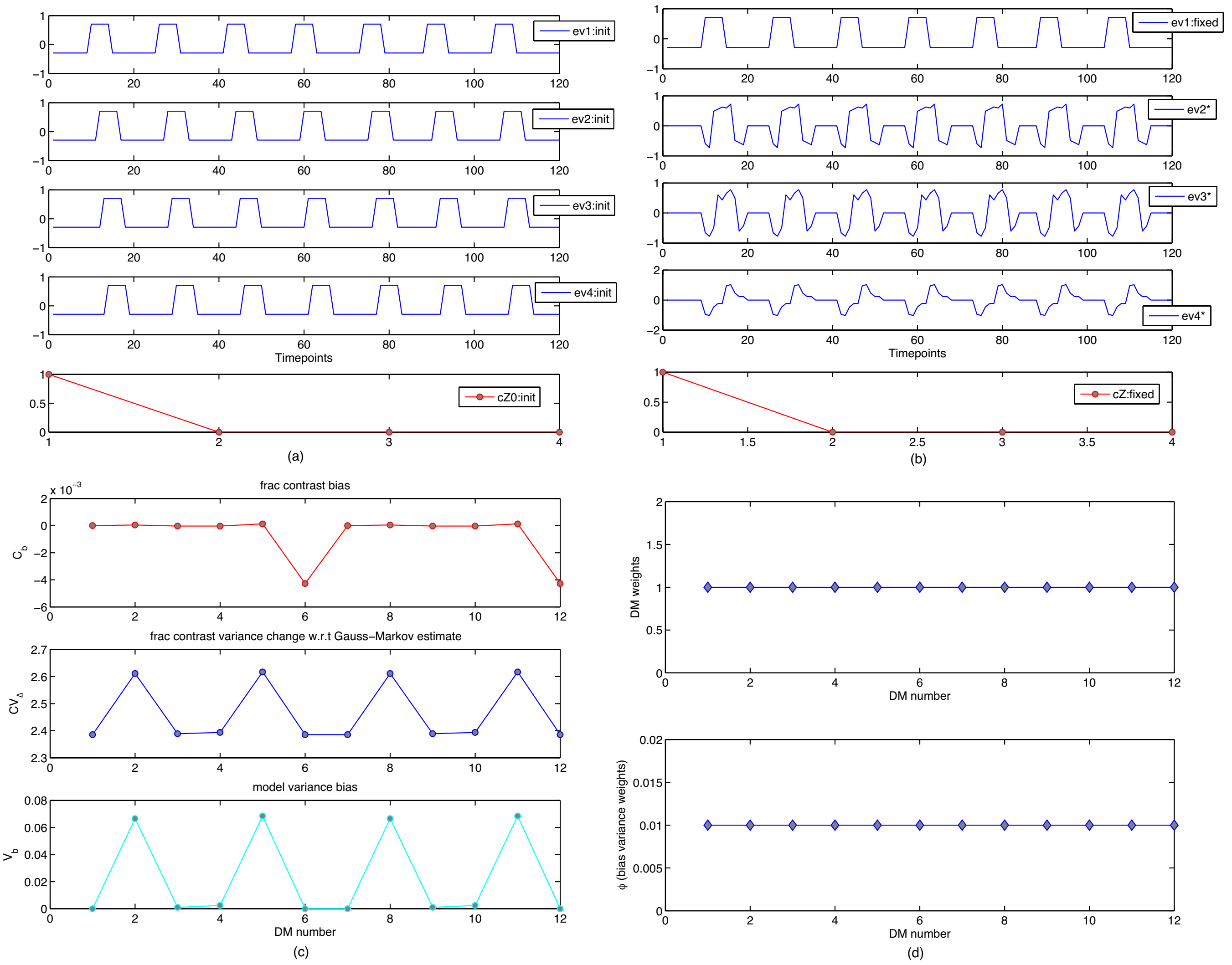


Fig. 12. Example 5: (a) Initial design matrix (DM) along with initialization of columns 2 to 4 of Z_0 using X_2 to X_4 respectively. The first column in Z was fixed to the basic block design EV, columns 2 to 4 in Z were left unconstrained and the contrast was fixed at [1;0;0;0] (b) Estimated optimal DM showing optimal columns 2 to 4 (c) Performance curves showing the fractional contrast bias C_b, contrast variance change w.r.t Gauss-Markov estimate CV_Δ and model variance bias V_b (d) In this example $w_i = 1$ and ϕ_i = 0.01 indicating a predominant weight to bias term during optimization.

subject data.

ADIS produced a total of 16 components. Figure 17 shows associated timecourses for some minimum entropy spatial maps. The linear drift was found to be present globally throughout the brain superimposed on the various response profiles. Thus an example of a potential design matrix would be the one such as shown in Figure 20 with the first column a ramped step change modeling the infusion response and the second column a covariate of no interest as the linear drift. The contrast of interest to the investigator will be $[1;0]$. Clearly the data contains multiple response profiles differing in mainly their rise from baseline. There could be many more variations possible in general datasets. In the study of pharmacological fMRI responses, the exact response of the brain is not easily modeled because of pharmacokinetics and pharmacodynamics differences between the circulating system and the brain. Hence, it is feasible to expect different responses across brain structures. It seems appropriate to use a "ramp" model with potential delays to accomodate responses in different brain structures.

3) Optimal DM computation: Subsequent to this data "inspection" stage we carried out an analysis to compute the optimal design matrix. Two main EVs were proposed for this analysis as shown in 20. The EV1 captures an "infusion response" while EV2 represents pure "linear drift". For the purposes of optimization we proposed 723 potential DMs with the intention of capturing delays upto 180 timepoints. Let X_0 be the DM shown in 20. Define

$$\begin{aligned} X_i(:,1) &= X_0(:,1) \text{ shifted to the right by } i \text{ timepoints} \\ X_i(:,2) &= X_0(:,2) \text{ for all } i \end{aligned} \tag{60}$$

These 723 DMs are as follows:

1) The first 180 DMs were $X_1, \ldots, X_{180}$ at $\frac{\beta_i}{\sigma_i} = [1; 0.5]$ and $c_{X_i} = [1; 0]$
2) The next 180 DMs were $X_1, \ldots, X_{180}$ at $\frac{\beta_i}{\sigma_i} = [-1; 0.5]$ and $c_{X_i} = [1; 0]$

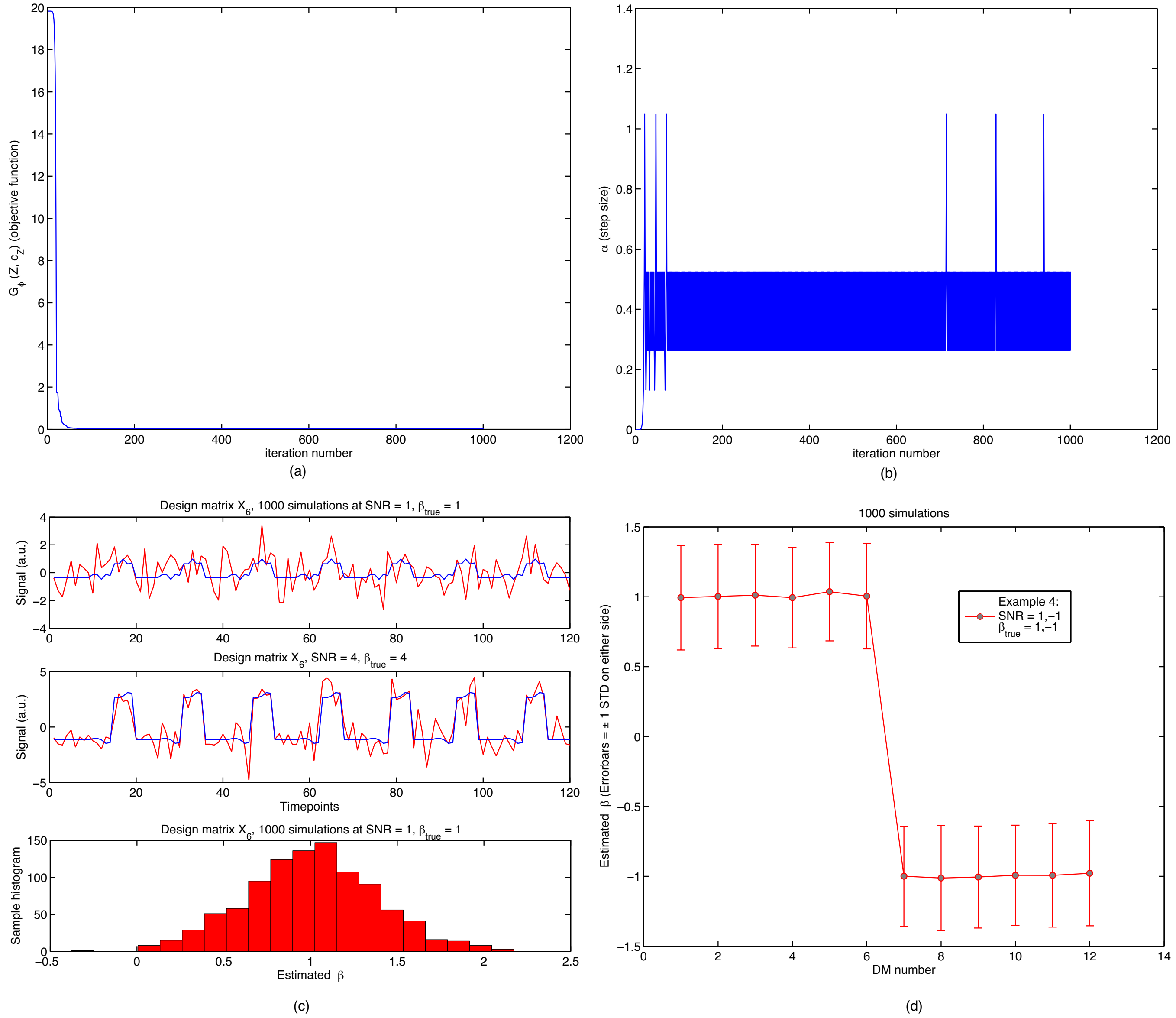


Fig. 13. Example 5: (a) Figure showing the evolution of objective function values $G_\phi(Z, c_Z)$ over algorithm iterations. Notice how the function value stabilizes as convergence is reached (b) Figure showing the variation in the step size α over algorithm iterations. Step size controlling parameter θ in Algorithm 1 was set to $\theta = 2$ (c) , (d) For each design matrix (DM) entered into optimization, 1000 simulated data-sets were generated at SNR $\frac{\beta_i}{\sigma_i}$. A GLM analysis was run on each of these data-sets using the optimized DM. Figure (c) shows an example of simulated data for DM X_6 at SNR $\frac{\beta_6}{\sigma_6}$ and the GLM fit using the optimal DM. We also show a GLM fit at a higher SNR of 4 for illustration purposes. It also shows the distribution of $c_Z^T\hat{\gamma}$ over 1000 simulations. Figure (d) is a summary errorbar plot showing $\hat{E}(c_Z^T\hat{\gamma})$ over 1000 simulations for data generated from each DM. The error bars represent unit standard deviation of $c_Z^T\hat{\gamma}$ (**not** standard deviation of $\hat{E}(c_Z^T\hat{\gamma})$) to quantify the variance in estimation via simulation.

3) The next 180 DMs were $X_1, \ldots, X_{180}$ at $\frac{\beta_i}{\sigma_i} = [1; -0.5]$ and $c_{X_i} = [1; 0]$
4) The next 180 DMs were X_1, $\ldots$, X_{180} at $\frac{\beta_i}{\sigma_i} = [-1; -0.5]$ and $c_{X_i} = [1; 0]$
5) DM 721 was X_0 at $\frac{\beta_i}{\sigma_i} = [0; 1]$ and $c_{X_i} = [1; 0]$
6) DM 722 was X_0 at $\frac{\beta_i}{\sigma_i} = [0; -1]$ and $c_{X_i} = [1; 0]$
7) DM 723 was X_0 at $\frac{\beta_i}{\sigma_i} = [0; 0]$ and $c_{X_i} = [1; 0]$

(1) to (4) above instruct the optimization process to be sensitive to both "activation" and "deactivation" with both "positive" and "negative" linear drift. We also explicitly instruct the optimization process to match pure "positive" or "negative" linear drift to an "infusion response" of size 0 using (5) and (6) above. (7) instructs the optimization to match "infusion response" to size 0 in the absence of any signal (either infusion response or linear drift). We choose a conservative SNR of 1 for the signal of interest and a linear drift amplitude 50% that of the main signal. The computed optimal DM and its performance are shown in Figures 18-19

For the optimization, we fixed the first two columns of Z to the EVs shown in 20. The contrast was also fixed to be $[1; \ldots; 0]$ corresponding to the main infusion response.

The purpose of running model free analysis was to demonstrate that a significant variation in the signal of interest is often present in real data. Our potential model set is not based

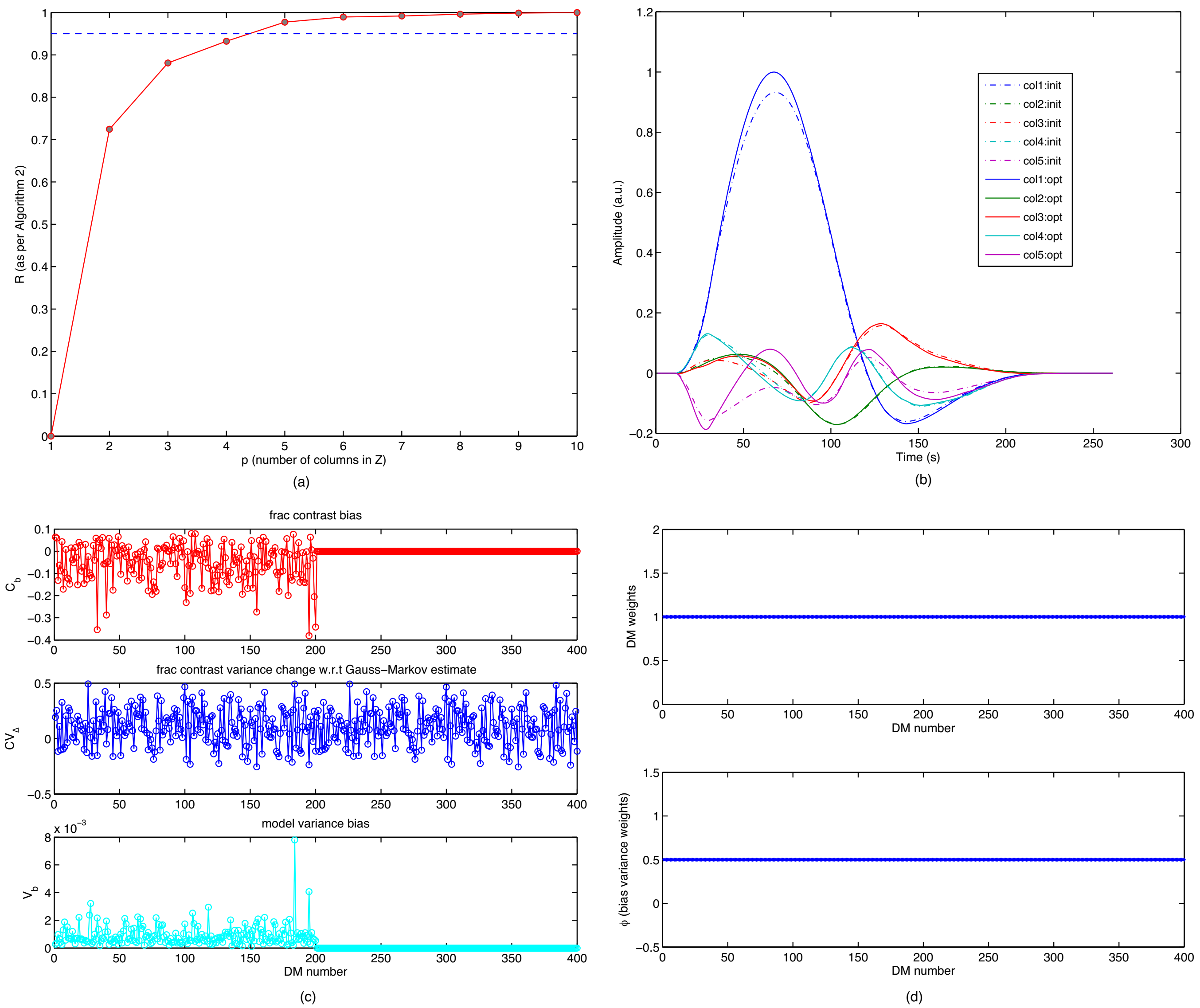


Fig. 15. Example 6: (a) R versus p curve for determining the optimal number of columns in Z. Using a cutoff of $R_c = 0.95$ the optimal number of columns in Z was determined to be $p_{opt} = 5$. (b) The 5 columns of Z were left unconstrained and the contrast was fixed at $[1;0;0;0]$. The automatic initialization strategy described before was used to initialize the 5 columns of Z_0 (dotted lines). The optimized columns are shown in the same figure using solid lines. (c) Performance curves showing the fractional contrast bias C_b, contrast variance change w.r.t Gauss-Markov estimate $CV_\triangle$ and model variance bias V_b (d) In this example $w_i = 1$ and ϕ_i = 0.5 indicating an equal weight to bias and variance terms during optimization.

on the results of model free analysis. For instance model free analysis detects only 6 examples of infusion responses shown in Figure 17 yet our proposed model set contains 723 models which is much larger than 6. The rationale for this selection is that the set of 723 models is large enough such that it will contain the infusion responses of ”interest”. The selection of this model set is also supported by studies of pharmacological fMRI (phMRI) responses.

However, we would like to caution the reader that when using model free analysis as a data inspection step in a group analysis it is important to add a cross-validation step to avoid ”circular analysis” (see discussion). Since group analysis and cross-validation are tangential to the purpose of the present case study namely that of illustrating computation of optimal design matrices, we have skipped the cross-validation step here for simplicity.

4) GLM analysis: Subsequent to the computation of optimal DM, two GLM analyses were carried out using FSL's tool FEAT: (1) using an optimized design matrix (2) using a design matrix containing 3 EVs. The 2 EVs shown in 20 and an additional EV representing the temporal derivative of the ”infusion EV”.

VI. Results

Figures 31 - 36 illustrate the results of validation tests for Algorithm 1. It was found that the maximum difference in objective function using Algorithm 1 and the more sophisticated optimization solver was of the order of 10^{-5} (see Table II).

A. Case studies 1-6

In Examples 1-4, the size of the optimal DM was chosen to be $p = 3$. The weights w_i were chosen to be 1 for all i to

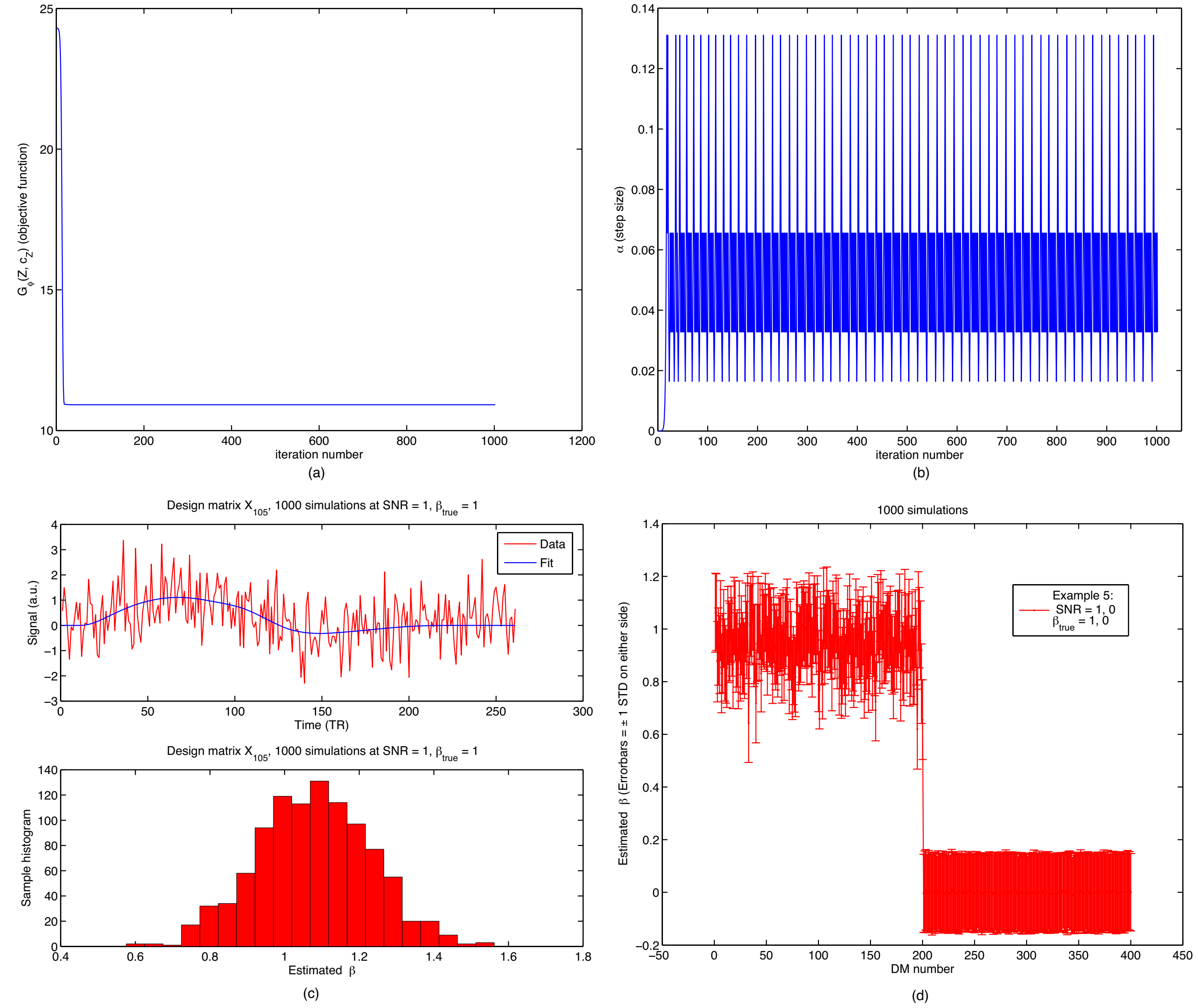


Fig. 16. Example 6: (a) Figure showing the evolution of objective function values $G_\phi(Z, c_Z)$ over algorithm iterations. Notice how the function value stabilizes as convergence is reached (b) Figure showing the variation in the step size α over algorithm iterations. Step size controlling parameter θ in Algorithm 1 was set to $\theta = 2$ (c) , (d) For each design matrix (DM) entered into optimization, 1000 simulated data-sets were generated at SNR $\frac{\beta_i}{\sigma_i}$. A GLM analysis was run on each of these data-sets using the optimized DM. Figure (c) shows an example of simulated data for DM X_{105} at SNR $\frac{\beta_{105}}{\sigma_{105}}$ and the GLM fit using the optimal DM. It also shows the distribution of $c_Z^T\hat{\gamma}$ over 1000 simulations. Figure (d) is a summary errorbar plot showing $\hat{E}(c_Z^T\hat{\gamma})$ over 1000 simulations for data generated from each DM. The error bars represent unit standard deviation of $c_Z^T\hat{\gamma}$ (**not** standard deviation of $\hat{E}(c_Z^T\hat{\gamma})$) to quantify the variance in estimation via simulation.

reflect equal likelihood of potential DMs in all Examples.

1) Example 1: Performance curves for Example 1 are shown in Figures 8 - 9. In this example ϕ_i was set to its default value of 0.5. Columns 1 and 2 of the Z were fixed during optimization and the unconstrained column in Z was initialized by drawing from a uniform distribution $U(0, 1)$.

It is seen that the contrast bias $|C_{bi}|$ is maintained at < 0.05 for $i = 1, \ldots, 36$. For $i = 37, \ldots, 50$, $|C_{bi}|$ increases to around 0.12 for $i = 50$. At the same time the contrast variance change w.r.t. Gauss-Markov estimate $CV_{\Delta i}$ decreases monotonically from 0.096 for $i = 1$ to 0.002 for $i = 20$ and becomes negative from $i = 21, \ldots, 50$ implying a lower variance than the Gauss-Markov estimator. The model variance bias V_{bi} is maintained at $< 2 \times 10^{-4}$ for all i.

2) Example 2: Performance curves for Example 2 are shown in Figures 27 - 28. In this example, ϕ_i was chosen automatically as described in section III-D. Columns 1 and 2 of the Z were fixed during optimization and the unconstrained column in Z was initialized at the optimal solution from Example 1.

It is seen from Figure 27 that lower weights ϕ_i were assigned for increasing i (approximately 20% lower for $i = 50$ compared to $i = 1$). With this choice of ϕ_i it was seen that the contrast bias $|C_{bi}|$ was reduced to < 0.01 for all $i < 36$. For $i = 37 \ldots, 50$, $|C_{bi}|$ increases to a maximum of 0.03 for $i = 50$. The value of $CV_{\Delta i}$ decreases monotonically from

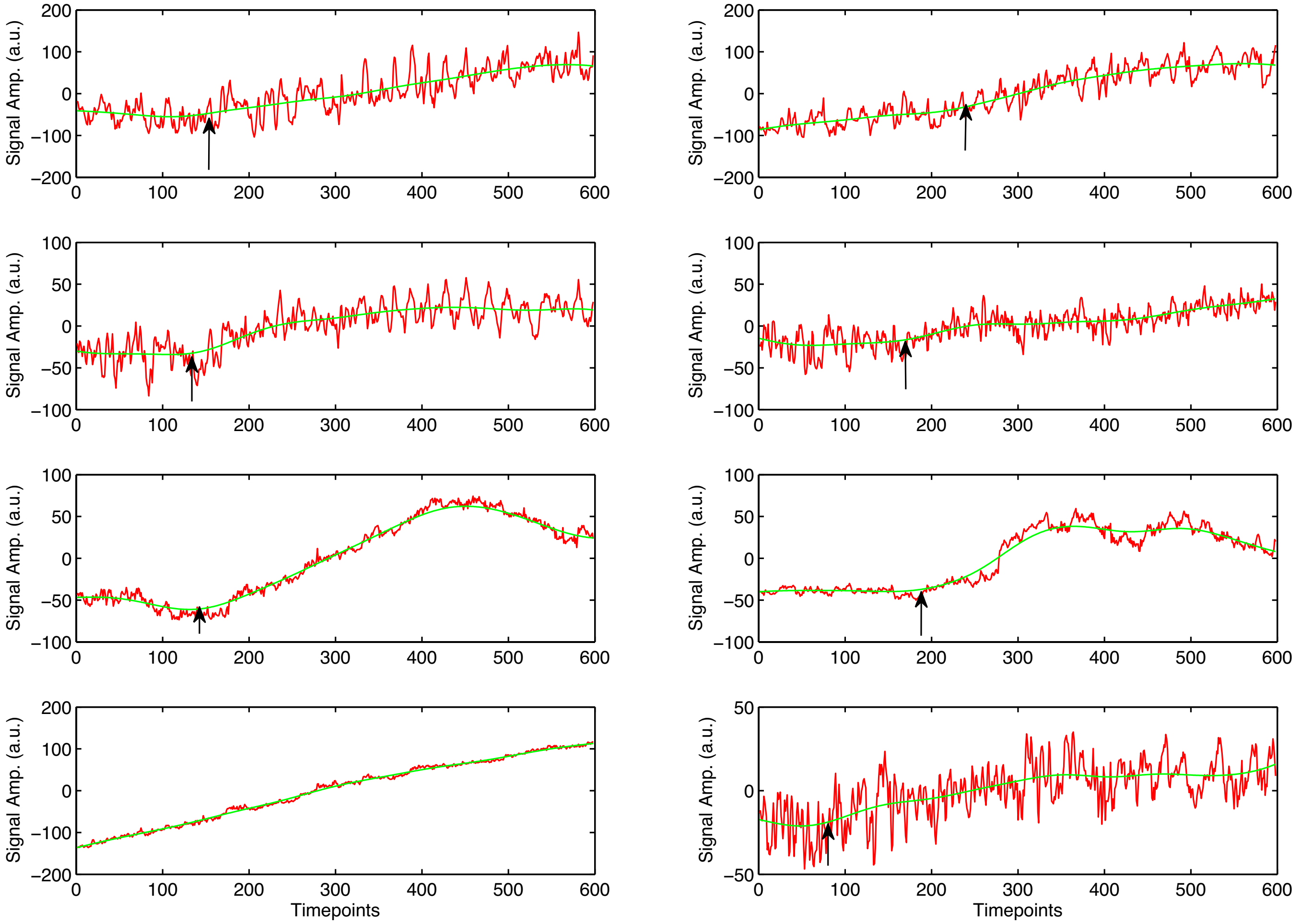


Fig. 17. Some infusion response profiles extracted by ADIS (arXiv:0902.4879v1) illustrating the variability of the signal of interest. The black arrow indicates the approximate timepoint at which the signal starts going up from baseline. Note that the individual infusion responses are potentially corrupted by the linear drift which was identified as a global component in the brain.

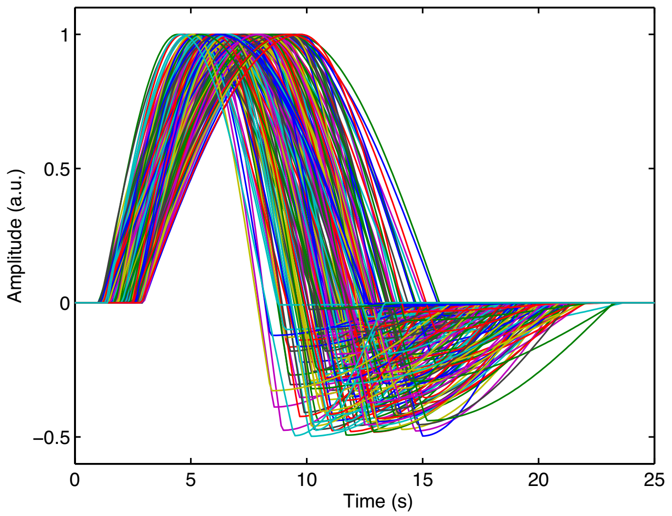


Fig. 14. 200 samples of HRF drawn from a 5 parameter half-cosine parameterization of Haemodynamic Response Function (HRF) that were used as input to the optimization process in Example 6.

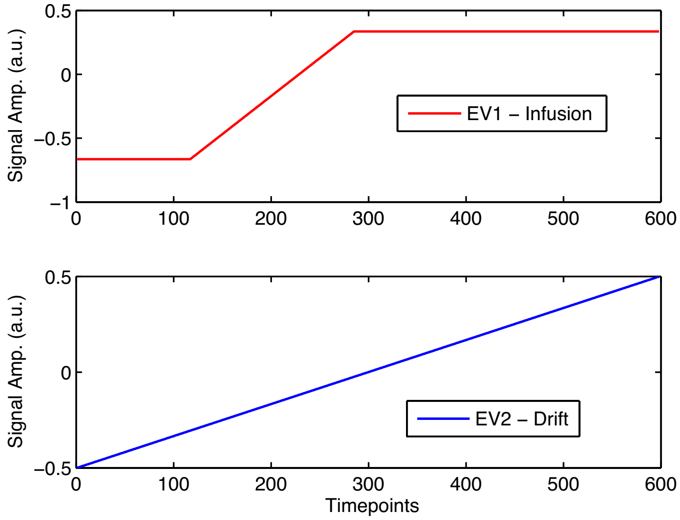


Fig. 20. Example infusion design matrix. The top figure shows a typical ”infusion response” and the bottom figure shows a confounding ”linear drift”.

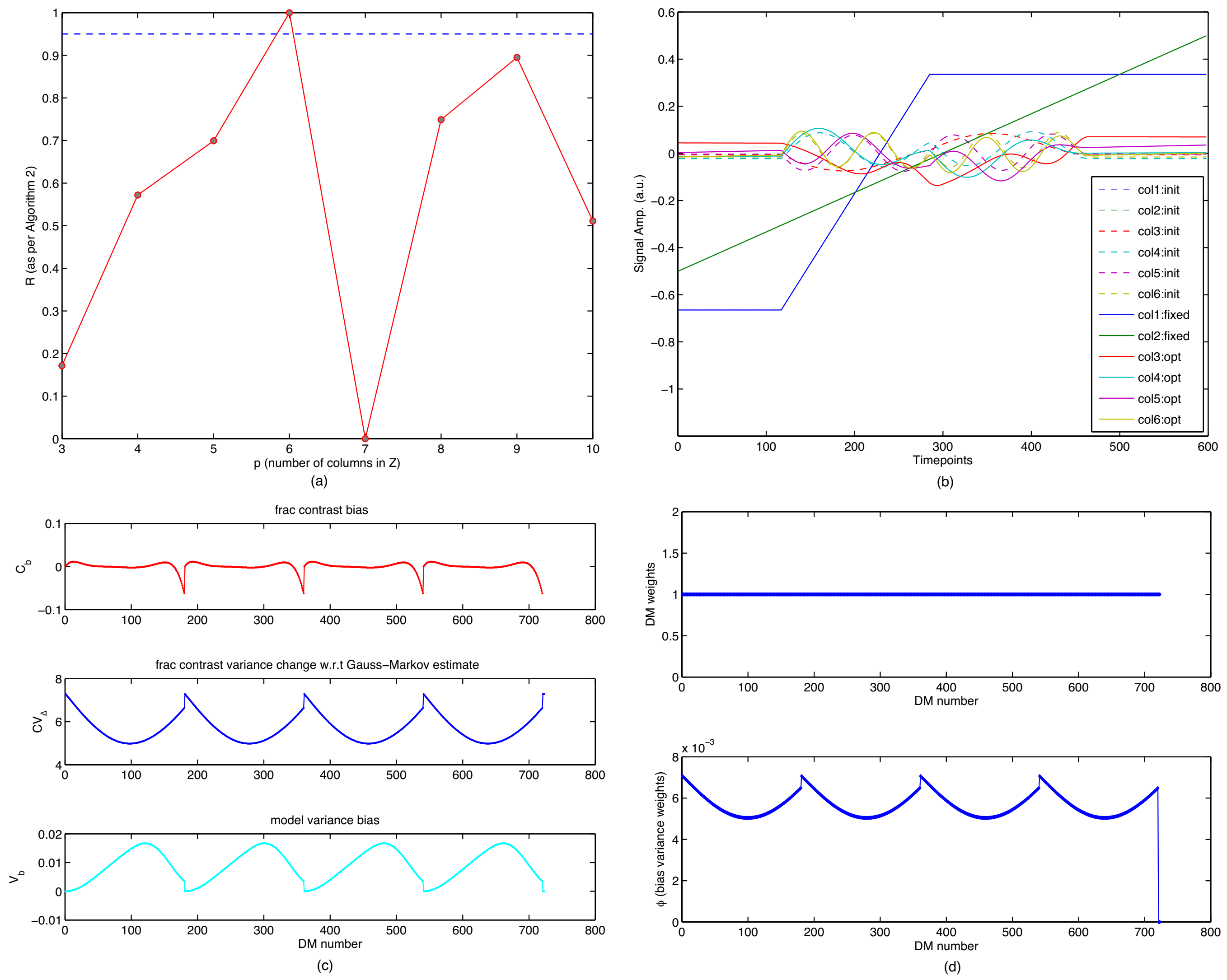


Fig. 18. fMRI case study: (a) R versus p curve for determining the optimal number of columns in Z. Using a cutoff of $R_c = 0.95$ the optimal number of columns in Z was determined to be $p_{opt} = 6$. (b) The first 2 columns of Z were constrained and the others were left unconstrained during optimization. The contrast was fixed at $[1; 0; 0; 0; 0; 0]$. The automatic initialization strategy described before was used to initialize the 6 columns of Z_0 (dotted lines). The optimized columns are shown in the same figure using solid lines. (c) Performance curves showing the fractional contrast bias C_b, contrast variance change w.r.t Gauss-Markov estimate CV_Δ and model variance bias V_b (d) In this example $w_i = 1$ and ϕ_i was chosen using the automatic initialization strategy described before.

0.345 for $i = 1$ to 0.07 for $i = 50$ implying an increased variance compared to Example 1. The model variance bias V_{bi} was maintained at $< 4 \times 10^{-3}$ for all i.

3) Example 3: Performance curves for Example 3 are shown in Figures 29 - 30. In this example, ϕ_i was chosen to be 0.1 for all i. Columns 1 and 2 of the Z were fixed during optimization and the unconstrained column in Z was initialized by drawing from a uniform distribution $U(0, 1)$.

It was found that $|C_{bi}|$ was maintained at < 0.02 for $i = 1, \ldots, 41$. For $i = 42, \ldots, 50$, $|C_{bi}|$ increases to 0.055 for $i = 50$. It was also found that $CV_{\Delta i}$ decreases monotonically from 0.195 for $i = 1$ to 0.003 for $i = 38$ and becomes negative from $i = 39, \ldots, 50$ indicating a lower variance than the Gauss-Markov estimator. The model variance bias V_{bi} was maintained $< 1 \times 10^{-3}$ for all i.

4) Example 4: Performance curves for Example 4 are shown in Figures 10 - 11. In this example, ϕ_i was chosen to be 0.01. Optimization was initialized using solution found in Example 1. The sample space of Z was left unconstrained but the set of potential DMs was augmented to explicitly instruct the optimization to enable detection of signals in the presence of confounds, as well as treat ”drift” signals as ”null” data.

It was found that $|C_{bi}|$ was reduced to < 0.02 for all i and $CV_{\Delta i}$ decreased monotonically from around 0.32 to around 0.05 for each of the four segments $i = 1, \ldots, 50$, $i = 51, \ldots, 100$, $i = 101, \ldots, 150$ and $i = 151, \ldots, 200$. The model variance bias V_{bi} was maintained $< 2 \times 10^{-3}$ for all i.

5) Example 5: Performance curves for Example 5 are shown in Figures 12 - 13. This example illustrated a block design experiment where the user wants to primarly control bias. ϕ_i was chosen to be 0.01 in this example indicating a preferential reduction of bias. The first column of the optimal DM was fixed to the primary block EV. The optimization

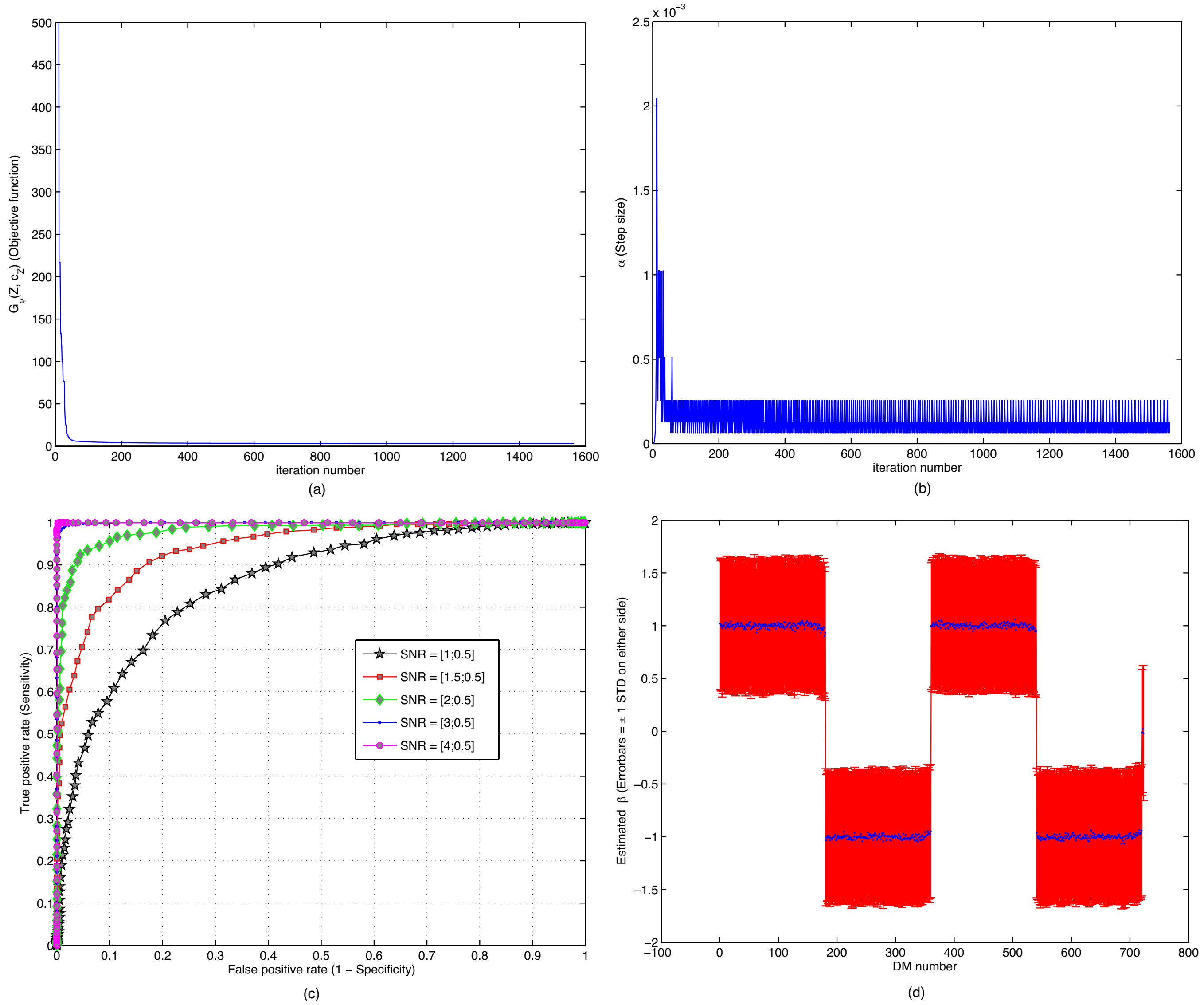


Fig. 19. fMRI case study: (a) Figure showing the evolution of objective function values $G_\phi(Z, c_Z)$ over algorithm iterations. Notice how the function value stabilizes as convergence is reached (b) Figure showing the variation in the step size α over algorithm iterations. Step size controlling parameter θ in Algorithm 1 was set to $\theta = 2$. For each design matrix (DM) entered into optimization, 1000 simulated data-sets were generated at SNR $\frac{\beta_i}{\sigma_i}$. A GLM analysis was run on each of these data-sets using the optimized DM. Figure (c) shows the ROC curve for data generated from the design matrix X_{180} at various SNR values. Figure (d) is a summary errorbar plot showing $\hat{E}(c_Z^T \hat{\gamma})$ over 1000 simulations for data generated from each DM. The error bars represent unit standard deviation of $c_Z^T \hat{\gamma}$ (**not** standard deviation of $\hat{E}(c_Z^T \hat{\gamma})$) to quantify the variance in estimation via simulation.

process was initialized using shifted versions of the primary EV.

It was found that the optimal DM reduced $|C_{bi}|$ to < 0.0042 for i while $CV_{\Delta i}$ was around 2.38 for $i = 1, 3, 4, 6, 7, 9, 10$ and 2.61 for $i = 2, 5, 8, 11$ indicating an increased variance relative to the Gauss-Markov estimator. The model variance bias V_{bi} was maintained $< 2.4 \times 10^{-3}$ for $i = 1, 3, 4, 6, 7, 8, 10$ and $< 6.7 \times 10^{-2}$ for $i = 2, 5, 8, 11$.

6) Example 6: Peformance curves for Example 6 are shown in Figures 15 - 16. This example illustrated the construction of a set of optimal HRF capturing functions that enable capture of locally variable HRF functions in the brain. The optimization process was explicitly indicated to match "null" data with signal size of "0" optimally using additional DMs from $i = 201, \ldots, 400$ at $\frac{\beta_i}{\sigma_i} = 0$. ϕ_i was set to its default value of 0.5. The size of optimal DM was found to be $p = 5$ using Algorithm 2. The 5 columns of optimal DM were left unconstrained during the optimization process and initialized using the procedure described in section III-B.

It was found that the mean absolute value of contrast bias was maintained at $|C_{bi}| < 0.075$ and the mean value of $CV_{\Delta i}$ was maintained at < 0.107 for all HRF shapes $i = 1, \ldots, 200$. For the "null" data $|C_{bi}| = 0$ and the mean $CV_{\Delta i}$ was maintained at < 0.106 for $i = 201, \ldots, 400$. The model variance bias V_{bi} was reduced to $< 8 \times 10^{-3}$ for all i.

B. fMRI case study

This case study deals with the optimal capture of signals for an fMRI infusion study. An initial model free exploration of data revealed the presence of multiple infusion profiles differing in their time to take off from baseline as illustrated in Figure 17. For the DM optimization, ϕ_i was initialized automatically using the strategy describe in section III-D. The first two columns of Z were fixed to the EVs described in Figure 20. The optimal number of columns in Z were estimated using Algorithm 2 to be $p = 6$. DM optimization was initialized automatically using the procedure described in section III-B. The set of potential DMs were augmented with additional DMs (total 723 potential DMs) to guarantee detection of ”positive” and ”negative” activation in the presence of confounding drift as well as to match both the ”drift” and ”null” data to a signal size ”0”.

Performance curves for the optimal DM are shown in Figures 18 - 19. It was found that the mean absolute value of $|C_{bi}|$ was reduced to 0.0057 and the mean value of $CV_{\Delta i}$ was 5.70 for $i = 1, \ldots, 720$ representing the ”non-null data”. For the ”null” data from $i = 721, \ldots, 723$, we found $|C_{bi}| \equiv 0$ and the mean $CV_{\Delta i}$ was 7.28. The model variance bias V_{bi} was reduced to $< 1.7 \times 10^{-2}$ for all i. Figure 19 shows the ROC curve for the detection of signal generated from X_{180} versus the ”null” data generated from a ”pure” drift signal at SNR $[0; 0.5]$ over 1000 simulations from each. We chose X_{180} because it is an extreme DM that produces the highest absolute contrast bias $|C_{bi}| \sim 0.063$. A simple decision rule based on the T statistic calculated from equation 34 was used for positive activation detection to generate the ROC curve. For a cutoff T-statistic t_c:

$$\text{Decide activation if: } T(\hat{\gamma}, \hat{\sigma_1}; \hat{Z}; \hat{c_Z}) \geq t_c \qquad (61)$$

$$\text{Decide null if: } T(\hat{\gamma}, \hat{\sigma_1}; \hat{Z}; \hat{c_Z}) < t_c \qquad (62)$$

where $\hat{\gamma}$ and $\hat{\sigma_1}$ are the parameter vector and residual standard deviation respectively estimated using the optimal DM $\hat{Z}$ and contrast $\hat{c_Z}$. It is found that for SNR of $\frac{\beta}{\sigma} = [2; 0.5]$, a sensitivity of 93.5% was obtained at a specificity of 94.2% and for SNR of $\frac{\beta}{\sigma} = [3; 0.5]$, a sensitivity of 98.8% was obtained at a specificity of 98.7%. The real fMRI data was found to have approximately an $SNR > 4$ for the primary infusion response indicating a high sensitivity and specificity of signal detection using the optimal DM $\hat{Z}$.

Figure 22 shows the t-stat image corresponding to the ”infusion response” overlaid onto the anatomical image (transformed into the ”fMRI” space) of the subject. For illustration purposes, we extracted raw timecourses from 6 sample points in the brain the details of which are shown in Table I. Figure 23 - 25 show the raw timecourses from real fMRI data with full and partial model fits obtained using the optimal DM $\hat{Z}$. For illustration purposes we also show the full and partial model fits obtained using a ”naive” DM such as that using the temporal derivative of the ”infusion response” as a covariate. Sample point 1 represents a canonical 0-delay infusion response, while sample point 2 represents a 200 timepoint delayed infusion response. Sample points 3 and 4 represent infusion responses with delays of around 100 timepoints while sample points 5 and 6 represent infusion responses with delays of around 150 timepoints. It was found that using the optimal DM $\hat{Z}$ it was possible to detect all infusion responses (delayed or not) in a robust and unbiased fashion.

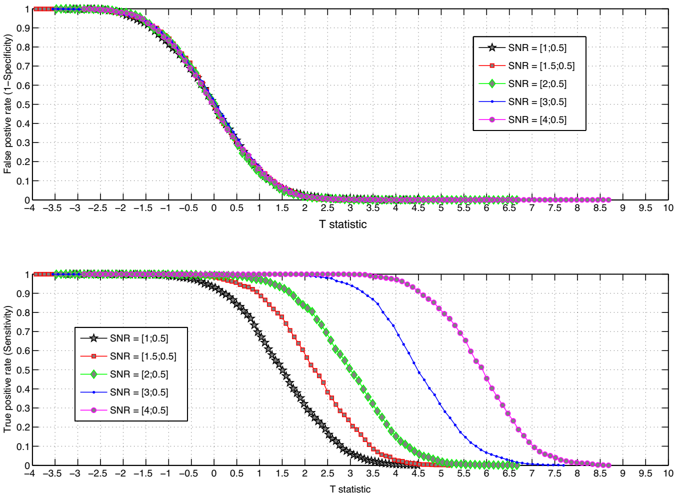


Fig. 21. Figure shows the variation of False positive rate (1 - Specificity) and True positive rate (Sensitivity) with the T statistic threshold for data generated as follows: (1) ”Activation” generated at various SNRs from an extreme DM - X_{180}, that produces the highest absolute contrast bias $|C_b| \sim 0.063$ of all 723 DMs used in optimizaton (2) ”Null” data generated from a ”pure drift” signal at SNR $[0; 0.5]$ which is a covariate of no interest and hence should be matched to a signal size of ”0”.

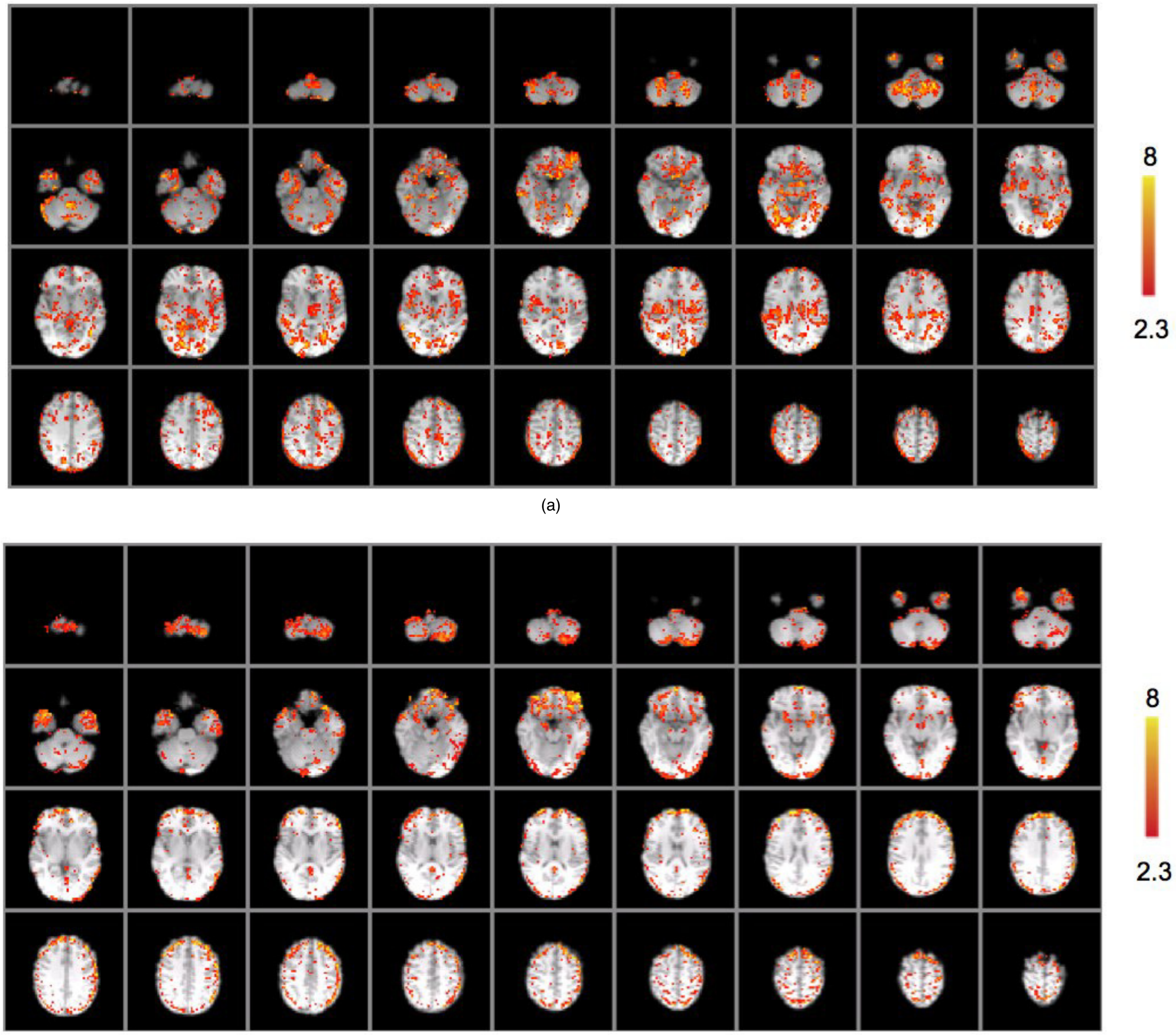
TABLE I
TABLE SHOWS 6 SAMPLE VOXELS, THEIR CO-ORDINATES AND THE GLM t-STAT VALUES FOR THE CONTRAST REPRESENTING THE ”SIZE” OF INFUSION EV USING THE OPTIMAL DESIGN MATRIX (OPT DM) AND THE DESIGN MATRIX WITH THE TEMPORAL DERIVATIVE (DER DM).

Sample	Coords	t-stat Opt DM	t-stat Der DM
1	[42,41,29]	4.84	8.56
2	[40,20,25]	4.06	-4.39
3	[42,28,24]	5.89	-0.86
4	[36,16,24]	4.45	-2.38
5	[18,26,24]	3.27	-1.67
6	[36,11,22]	4.35	-1.17

VII. DISCUSSION

The objectives of this paper were the development of a theoretical framework and a numerical algorithm to enable optimization of design matrices used in fMRI analyses. This optimization framework allows a user to optimize an objective function capturing the bias-variance decomposition over a set of potential design matrices. Within this very general framework it is possible to specify weights measuring the expected frequency of occurrence of various design matrices as well as preferentially control for bias or variance, if desired. The sample space for optimization is controlled via constraints on the columns of the optimal design matrix and the associated contrast.

We validated our numerical algorithm by comparing it with a more sophisticated optimization solver in two validation tests. We proposed a strategy for choosing the number of columns in the optimal design matrix as well as strategies for automatic selection of initial point for the optimization



(a)

(b)

Fig. 22. (a) fMRI data was analyzed using GLM with the optimal DM. Figure shows t-stat map for the contrast of interest $[1; 0; 0; 0; 0; 0]$ representing the ”infusion response”. The anatomical image for the sample subject was transformed to native space for t-map display purposes. The t-map was thresholded at an uncorrected threshold of 2.3. From the ROC curve shown in Figure 21, a specificity of $\sim 99\%$ and a sensitivity of $> 98\%$ is obtained at SNRs ≥ 3. Real fMRI data had an SNR > 4 and hence a cutoff of 2.3 on the t-statistic map performs well. (b) The same data from (a) is analyzed using a ”temporal derivative” based DM. This DM performs poorly and even detects significant ”positive” activation incorrectly as significant ”negative” activation indicating a more than 100% estimation bias. See Figures 23 - 25 for example timecourses.

algorithm and for choosing local bias-variance weightings ϕ_i based on user-specified objectives.

We then illustrated the application of the technique by considering 6 case studies. Our aim in these examples was to illustrate the degree of control that a user has in terms of controlling the optimal solution based on user specifications. The first 4 examples illustrated the variation of various control parameters in the algorithmic framework as applied to a phMRI study. Example 5 illustrated the application of the proposed technique to a block design case study. Example 6 addressed the important issue of locally variable HRF functions in fMRI data. The goal in example 6 was to come up with a set of HRF-modeling functions to capture a range of HRF shapes while maintaining optimality with respect to bias and variance of the primary contrast capturing the response amplitude of the underlying EV.

We examined how the optimized design matrix compared to an alternative in which a temporal derivative is added as an additional EV to capture variation in signal onset. Figure 26 compares the bias in parameter estimate at SNR = 1 over 1000 simulations for the temporal derivative approach as well as for the four optimized phMRI design matrices. A significant

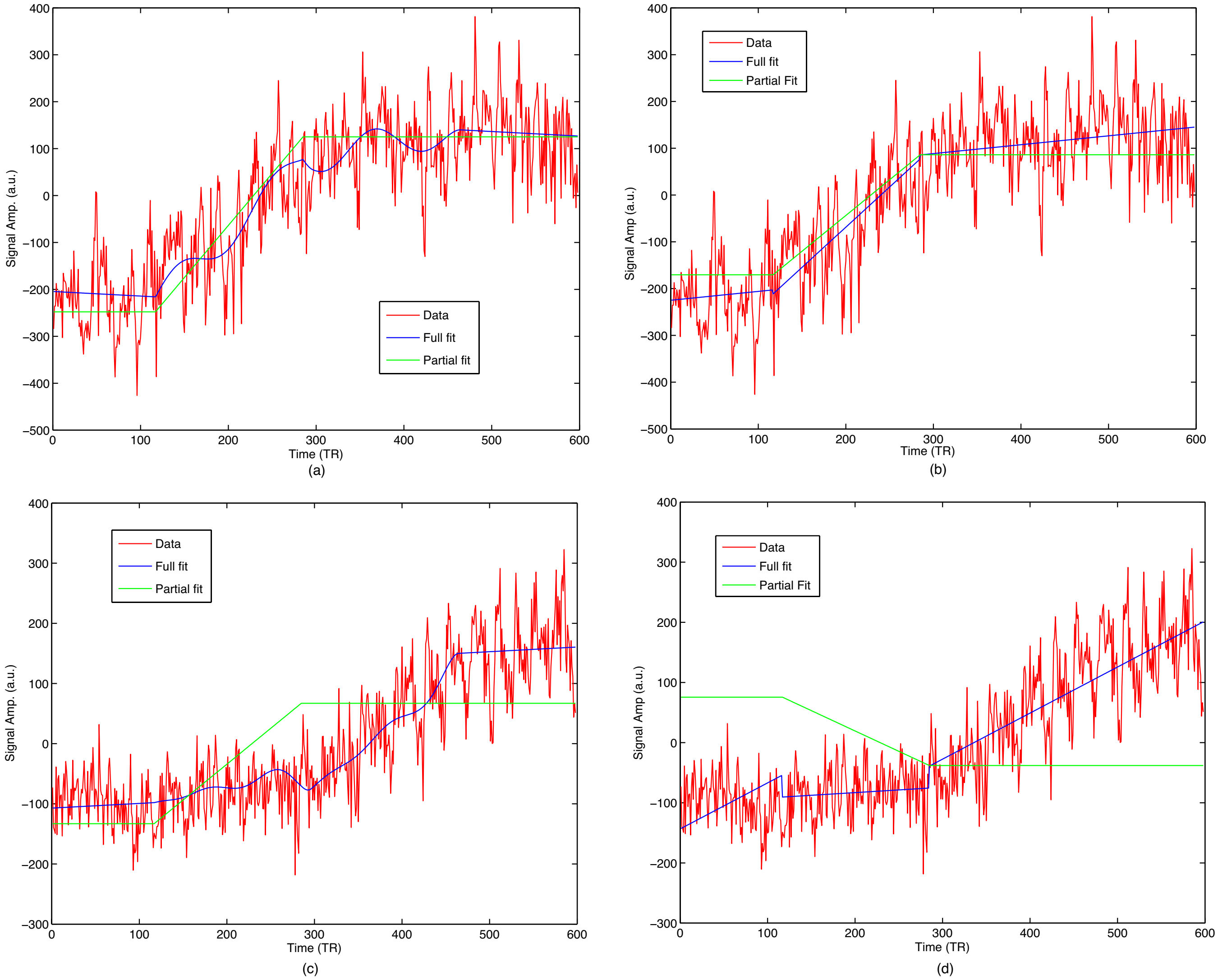


Fig. 23. fMRI data extracted from sample voxels 1 and 2. A GLM analysis was run using the optimal DM and a DM containing the 2 EVs shown in 20 and the temporal derivative of the 1st EV (i.e., the infusion ev). Figures show the full model fit and partial fit corresponding to the contrast of interest. Figures (a) and (c) are for the optimal DM while (b) and (d) are for the ”temporal derivative” based DM.

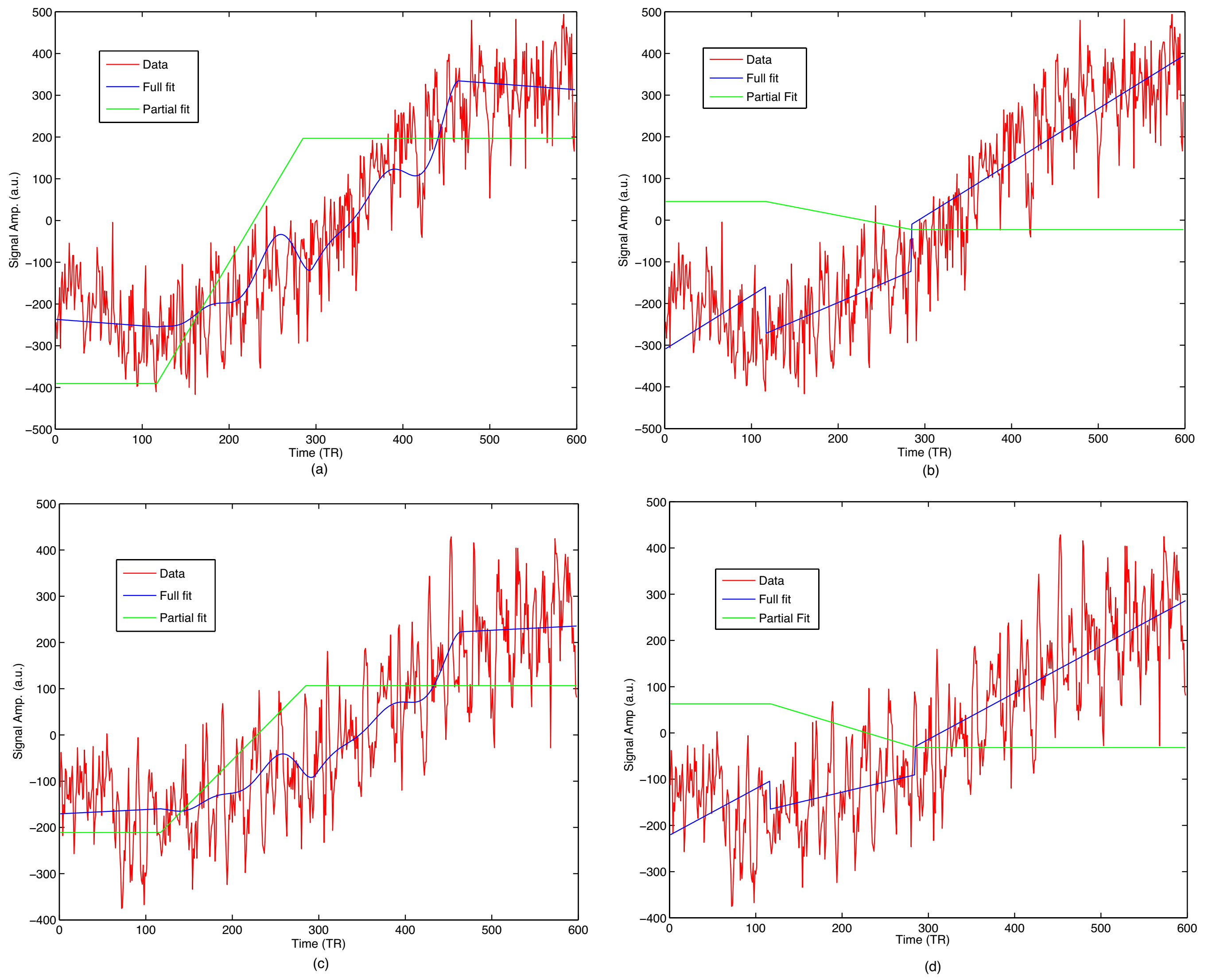
reduction in bias is observed when using the optimized design matrices in comparison to the temporal derivative approach, with Examples 2-4 performing best.

Finally, we applied the technique to a real infusion phMRI dataset. First, an ”optimal” design matrix was derived for this data. Next, we examined the bias-variance properties and generated ROC curves for the estimated design matrix. Finally, a GLM analysis was performed on the phMRI dataset using this design matrix. It was found that this design matrix achieves very high ($> 98\%$) sensitivity and specificity of signal detection over a range of signal variations in the data.

The use of a set of basis functions has been recommended before for capturing variability in fMRI signal shapes. For example, as relates to the capture of locally varying HRF shapes the work by FMRIB analysis group on FLOBS [9] is notable in that it constrains the basis set to have sensible HRF shapes. These approaches allow one to test the hypothesis about the presence or absence of signal by using F-statistics. However, limitations of such approaches include (1) the inability to measure the amplitude of the HRF signals and (2) The inability to combine amplitude measures from single-subject analyses into a group level analysis. Because the signal amplitude is never measured in these basis function approaches it is also not controlled for bias and variance. In contrast, the HRF-capturing functions developed in Example 6 illustrate how variable signal shapes can be captured using a single contrast while optimizing the bias and variance of the resulting signal amplitude estimates as per user defined objectives.

In [10], the authors address an important issue of mismodeling the fMRI signal by general linear model (GLM). They provide a method for detection of mismodeled regions by investigating GLM residuals. On the other hand, our work can be understood as complementary as it tries to generate the best design matrix for a problem at hand. One possible approach would be to use our method on the voxels flagged as mismodelled by [10]. This will be a topic for future research.

Inspection of residuals is a useful technique for detecting underfitting since this is when the residuals will not follow



(a)

(b)

(c)

(d)

Fig. 24. fMRI data extracted from sample voxels 3 and 4. A GLM analysis was run using the optimal DM and a DM containing the 2 EVs shown in 20 and the temporal derivative of the 1st EV (i.e., the infusion ev). Figures show the full model fit and partial fit corresponding to the contrast of interest. Figures (a) and (c) are for the optimal DM while (b) and (d) are for the ”temporal derivative” based DM.

an assumed null distribution. However if the fitted model is sufficiently complex the residuals can be ”small” and it may be difficult to detect mismodeling by only inspecting the residuals. We would like to point out that similar to [10] our approach also prefers to minimize residuals after model fit by including a residual related term in the objective function. However, there are two other terms that are related to bias and variance that prevent both underfitting and overfitting.

What is a ”true model”? In our opinion, ”true model” is only a concept that exists in the mind of the modeler. Unless the data has been generated by simulation, the ”true model” is always unknown. Therefore basing inferences on an ”assumed model” as if there is no model uncertainity will typically lead to a sense of falsely inflated precision and lack of robustness in the analysis. Instead of choosing an ”assumed model” we could choose a ”set of models”. In this case, our assumption would be that the ”true model” belongs to the chosen ”set of models”. Notice that this is a much weaker and more realistic assumption compared to a ”single model” assumption. It might also be much easier to specify a ”set of models” of which the true model is a subset instead of specifying a ”single model”. Our framework allows the capture of this ”model uncertainity” while optimizing the bias-variance tradeoff to enable robust inference.

Recent work by [11] cautions against the problem of ”circular analysis” in systems neuroscience. In simple terms ”circular analysis” refers to situations where ”analysis A” on data D is used to learn an ”assumption S” for ”analysis B” on the same data D. The problem with this approach is that assumption S is not independent of the data D and hence could distort the results of ”analysis B”. The degree of distortion in ”analysis B” will depend on the strength of dependence of S on the data D. An illustration of ”circular analysis” is when a model free analysis is run on a single subject fMRI data to select a ”single model” which is then used for GLM analysis of the same data. A simple way of avoiding ”circular analysis” is to

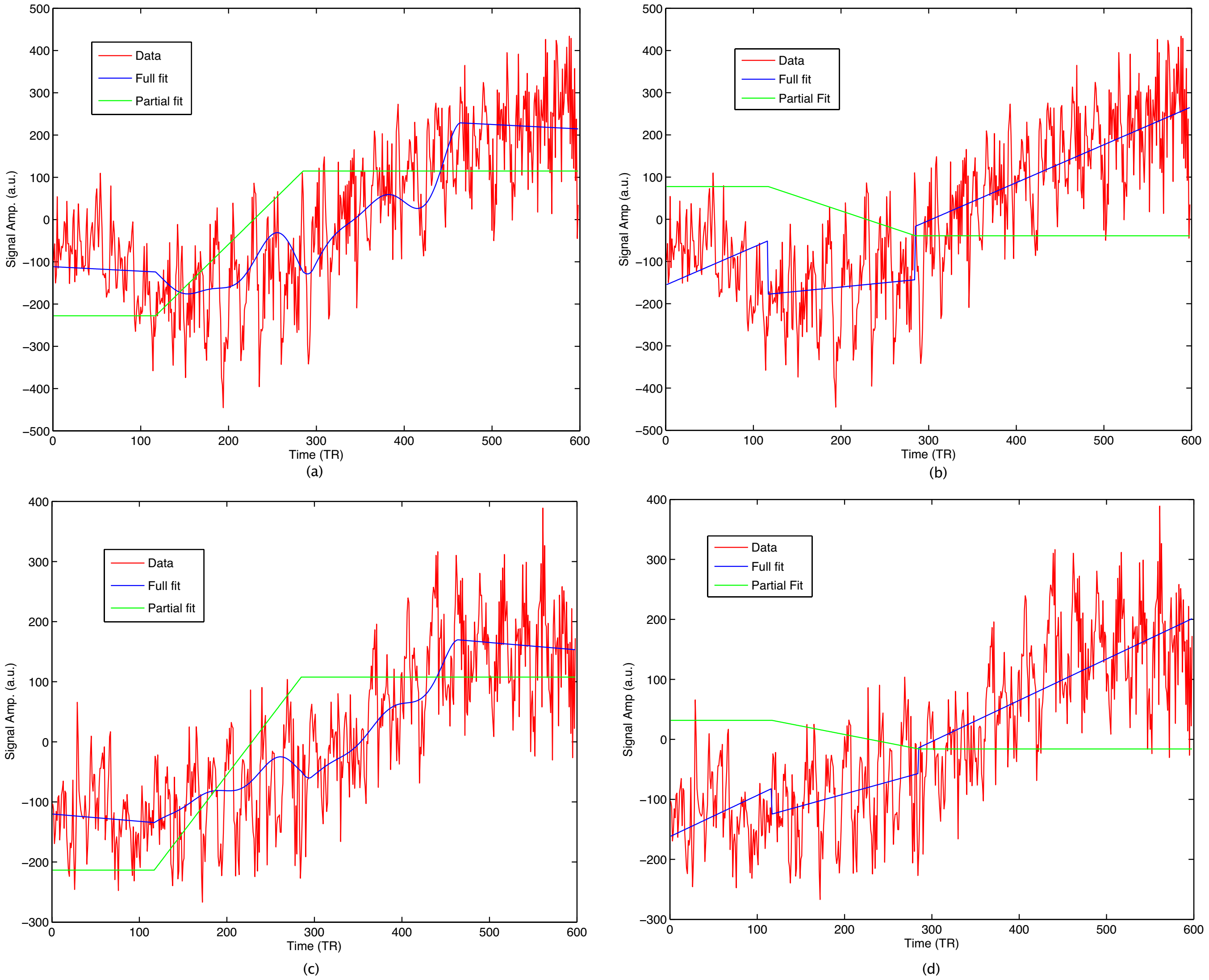


Fig. 25. fMRI data extracted from sample voxels 5 and 6. A GLM analysis was run using the optimal DM and a DM containing the 2 EVs shown in 20 and the temporal derivative of the 1st EV (i.e., the infusion ev). Figures show the full model fit and partial fit corresponding to the contrast of interest. Figures (a) and (c) are for the optimal DM while (b) and (d) are for the ”temporal derivative” based DM.

use assumption S (learned from ”analysis A” on data D) for ”analysis B” on new data D_{new}. This important issue must be kept in mind when choosing a ”set of models” for optimization of Z and c_Z. If group analysis is of interest then we would want our ”set of models” to contain the ”true group response” and not be biased by single subject data. We discuss below the main scenarios of interest:

1) **Specification of a ”set of models” apriori**
   In this case, the modeler specifies a ”set of models” that are expected to contain the ”true group response” based on prior knowledge of the problem at hand. An example is the use of a 5 parameter half-cosine parameterization to model the human brain’s HRF function. In this case the problem of ”circular analysis” does not occur.
2) **Analyzing new fMRI experiments**
   We illustrate this case with an example. Suppose we perform an fMRI experiment and we have no idea what the brain response will look like. There is a strong argument in favor of model free exploratory analysis in this case. Model free analysis can inform us what models to propose for optimization of Z and c_Z. Suppose we have 12 subjects in our study. Clearly, we do not want to bias our ”model set” by basing it on single subject data from any one of the 12 subjects. How should we avoid the problem of ”circular analysis” in this case? For such a situation, we would suggest using cross-validation where we split the data into a ”training” set and a ”test” set. We could use the ”training data” for selection of ”set of models” for optimization of Z and c_Z and apply the estimated model to the ”test” set. Inferences for group analysis could then be based on this ”test” set. This way the data used for estimation of model is independent of the data used for making inferences thereby eliminating the problem of ”circular analysis”.

Another problem that occurs especially in the context of fMRI is that of multiple comparisons. If Z is forced to contain regressors that model the "null" part of the data then under the null hypothesis the T statistic defined in 34 will have a central t-distribution 35. Hence standard techniques for addressing the multiple comparisons problem (see [12]) can be used for thresholding the statistical maps generated using Z and c_Z. When regressors modeling the "null" part of the data are not contained in Z (e.g. entire Z is optimized) then under the null hypothesis T will have a non-standard distribution. In this case, non-parametric procedures (see [13], [14], [15]) for multiple comparisons such as permutation tests based on the maximum T statistic or maximum cluster size can be used.

In this paper we considered an inverse problem. Instead of proposing a statistical estimator and studying its bias/variance properties, we define an objective function that captures the bias-variance decomposition over the set of potential signal shapes in the data and then explicitly optimize this objective function for both a design matrix and a contrast. This results in an optimized design matrix and contrast that automatically captures the amplitude of signals of interest. The resulting PE values can be easily carried over for a group level analysis.

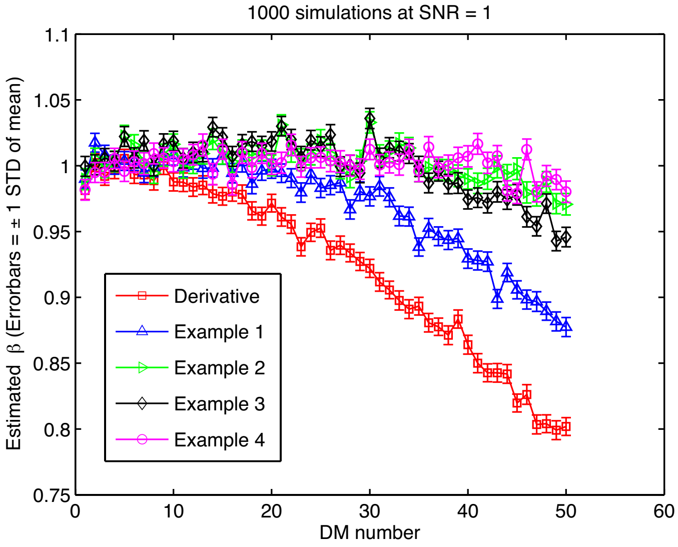


Fig. 26. For each design matrix (DM) X_i from Examples 1, 2, 3 and 4 entered into optimization, 1000 simulated data-sets were generated at SNR $\frac{\beta_i}{\sigma_i}$. A GLM analysis was run on each of these data-sets using the optimized DMs for Example 1, Example 2, Example 3, Example 4 as well as a DM X_D such that the first two columns of X_D are the same as the first two columns in $\hat{Z}$ from Example 1 but the 3rd column is the temporal derivative of $X_D(:, 1)$. Figure above shows a summary errorbar plot showing $\hat{E}(c_Z^T \hat{\gamma})$ over 1000 simulations for data generated from each DM and analyzed via GLM using optimized DM's from Example 1, 2, 3, 4 as well as using X_D. The errorbars represent the standard deviation of $\hat{E}(c_Z^T \hat{\gamma})$ so that bias can be statistically compared across the four cases.

VIII. Conclusion

We developed a theoretical framework to enable calculation of optimal design matrices for fMRI analyses that simultaneously enables detection of multiple signal responses as well as controlling for bias and variance in the GLM estimation. This is achieved by optimizing for the contrasts and basis functions of the design matrix to minimize the bias-variance decomposition over a set of design matrices capturing anticipated variability in the data. Although the technique was developed with motivation from fMRI, its development is quite general and appears to be applicable to a variety of problems from many other disciplines.

IX. Appendix

A. Non-central χ^2 distribution

If $X_i, i = 1, \ldots, \nu$ are independent random variables with a Gaussian distribution i.e., $X_i \sim N(\mu_i, \sigma_i^2)$ then the random variable $Z = \sum_{i=1}^{\nu}(\frac{X_i}{\sigma_i})^2$ is distributed according to the non-central χ^2 distribution. This distribution has two free parameters, the degrees of freedom ν and the non-centrality parameter Δ which is defined as $\Delta = \sum_{i=1}^{\nu}(\frac{\mu_i}{\sigma_i})^2$. We write $Z \sim \chi^2(\nu, \Delta)$ to indicate the Z has a non-central χ^2 distribution with degrees of freedom ν and non-centrality parameter Δ. We would like to point out that some authors define the non-centrality parameter as one half of what we define as the non-centrality parameter i.e., $\Delta_2 = \frac{1}{2}\sum_{i=1}^{\nu}(\frac{\mu_i}{\sigma_i})^2 = \frac{1}{2}\Delta$. Both parameterizations are completely equivalent as long as one keeps track of which parameterization is currently in use. Using our parameterization $E(Z) = \nu + \Delta$ and $Var(Z) = 2\nu + 4\Delta$ while using the alternative parameterization $E(Z) = \nu + 2\,\Delta_2$ and $Var(Z) = 2\nu + 8\Delta_2$. See [2] for further details.

B. Propositions and proofs

Suppose $y \sim N(X\beta, \sigma^2 I_n)$ where $X \in \mathbf{R}^{n\times q}$, $\beta \in \mathbf{R}^q$, $y \in \mathbf{R}^n$. Suppose matrix $Z \in \mathbf{R}^{n\times p}$ has full column rank (i.e. rank(Z) = p). Consider the equations for GLM estimation using matrix Z on data y:

$$\hat{\gamma} = (Z^T Z)^{-1} Z^T y \tag{63}$$

and

$$\hat{\text{Cov}}(\hat{\gamma}) = \hat{\sigma}_1^2\,(Z^T Z)^{-1} \tag{64}$$

where

$$\hat{\sigma}_1^2 = \frac{(y - Z\hat{\gamma})^T (y - Z\hat{\gamma})}{n - p} \tag{65}$$

The following propositions are in relation to this setup. To simplify notation, we will use a bold face zero $\mathbf{0}$ to indicate a matrix or vector of all zeros the size of which should be clear from context. A scalar zero will be indicated by 0.

Proposition 9.1: The following results hold:

$$E(\hat{\gamma}) = (Z^T Z)^{-1} Z^T X\beta \tag{66}$$

$$\hat{\gamma} \sim N\left((Z^T Z)^{-1} Z^T X\beta, \sigma^2 (Z^T Z)^{-1}\right) \tag{67}$$

$$(n-p)\frac{\hat{\sigma}_1^2}{\sigma^2} \sim \chi^2(n-p, \Delta) \tag{68}$$

where $\chi^2(n-p, \Delta)$ is a non-central χ^2 distribution with degrees of freedom $(n-p)$ and non-centrality parameter Δ (see IX-A). The non-centrality parameter Δ is given by:

$$\Delta = \frac{\beta^T X^T P_Z X\beta}{\sigma^2} \tag{69}$$

where

$$P_Z = I_n - Z(Z^T Z)^{-1} Z^T \tag{70}$$

Proof: Equation 66 readily follows by taking expectations on both sides of equation 63. Similarly 67 follows from $y \sim N(X\beta, \sigma^2 I_n)$ and the definition of $\hat{\gamma}$. Next we prove 68. Let $Z = Q_Z R_Z$ be the Q-R factorization of matrix Z. Since Z has full column rank, R_Z is non-singular and Q_Z is a $n \times p$ orthogonal matrix i.e., $Q_Z^T Q_Z = I_p$. Let $\tilde{Q}_Z$ be the $n \times (n-p)$ matrix that is the orthogonal complement of Q_Z i.e, $\tilde{Q}_Z^T \tilde{Q}_Z = I_{n-p}$ and $\tilde{Q}_Z^T Q_Z = \mathbf{0}$. Then $Q = [Q_Z, \tilde{Q}_Z]$ is a $n \times n$ orthogonal square matrix since $Q^T Q = I_n$. Since Q is square it is also true that $QQ^T = I_n$. This implies

$$Q_Z Q_Z^T + \tilde{Q}_Z \tilde{Q}_Z^T = I_n \tag{71}$$

Let P_Z be the orthogonal projector to the columns of Z. Using 63, the Q-R factorization of Z and 71 we can write:

$$P_Z = \left[I_n - Z(Z^T Z)^{-1} Z^T\right] = \left[I_n - Q_Z Q_Z^T\right] = \tilde{Q}_Z \tilde{Q}_Z^T \tag{72}$$

and therefore

$$y - Z\hat{\gamma} = \left[I_n - Z(Z^T Z)^{-1} Z^T\right] y = \tilde{Q}_Z \tilde{Q}_Z^T y \tag{73}$$

From 73 and 65 it follows that:

$$\hat{\sigma}_1^2 = \frac{\left[\tilde{Q}_Z \tilde{Q}_Z^T y\right]^T \left[\tilde{Q}_Z \tilde{Q}_Z^T y\right]}{n-p} \tag{74}$$

Simplifying and using the orthogonality of $\tilde{Q}_Z$ we get:

$$\hat{\sigma}_1^2 = \frac{y^T \tilde{Q}_Z\, \tilde{Q}_Z^T y}{n-p} = \frac{(\tilde{Q}_Z^T y)^T\, (\tilde{Q}_Z^T y)}{n-p} \tag{75}$$

Therfore

$$\frac{(n-p)\,\hat{\sigma}_1^2}{\sigma^2} = \left(\frac{\tilde{Q}_Z^T y}{\sigma}\right)^T \left(\frac{\tilde{Q}_Z^T y}{\sigma}\right) \tag{76}$$

Given that $y \sim N(X\beta, \sigma^2 I_n)$ we can write:

$$\frac{y}{\sigma} \sim N\left(\frac{X\beta}{\sigma}, I_n\right) \tag{77}$$

Thus it follows that:

$$\frac{\tilde{Q}_Z^T y}{\sigma} \sim N\left(\frac{\tilde{Q}_Z^T X\beta}{\sigma}, \tilde{Q}_Z^T \tilde{Q}_Z\right) \tag{78}$$

Using the orthogonality of $\tilde{Q}_Z$ this can be written as:

$$\frac{\tilde{Q}_Z^T y}{\sigma} \sim N\left(\frac{\tilde{Q}_Z^T X\beta}{\sigma}, I_{n-p}\right) \tag{79}$$

Thus elements of the $(n-p) \times 1$ vector $\frac{\tilde{Q}_Z^T y}{\sigma}$ are independent Gaussian random variables with the mean contained in the corresponding element of the vector $\frac{\tilde{Q}_Z^T X\beta}{\sigma}$ and unit variance. Hence from 76 and the definition of non-central χ^2 distribution IX-A it follows that:

$$\frac{(n-p)\,\hat{\sigma}_1^2}{\sigma^2} = \left(\frac{\tilde{Q}_Z^T y}{\sigma}\right)^T \left(\frac{\tilde{Q}_Z^T y}{\sigma}\right) \sim \chi^2(n-p, \Delta) \tag{80}$$

with the non-centrality parameter given by:

$$\Delta = \left(\frac{\tilde{Q}_Z^T X\beta}{\sigma}\right)^T \left(\frac{\tilde{Q}_Z^T X\beta}{\sigma}\right) = \frac{\beta^T X^T\, \tilde{Q}_Z \tilde{Q}_Z^T\, X\beta}{\sigma^2} \tag{81}$$

Using 72 this can be written as:

$$\Delta = \frac{\beta^T X^T\, \tilde{Q}_Z \tilde{Q}_Z^T\, X\beta}{\sigma^2} = \frac{\beta^T X^T\, P_Z\, X\beta}{\sigma^2} \tag{82}$$

This proves the result. ▮

■

Proposition 9.2: $\hat{\gamma}$ and $\hat{\sigma}_1^2$ are independent random variables.

Proof: We will use the Q-R factorization of Z from the proof of Proposition 9.1. Substituting $Z = Q_Z R_Z$ into 63 we get:

$$\hat{\gamma} = (Z^T Z)^{-1} Z^T y = R_Z^{-1} \, Q_Z^T y \tag{83}$$

As shown in 75

$$\hat{\sigma}_1^2 = \frac{(\tilde{Q}_Z^T y)^T \, (\tilde{Q}_Z^T y)}{n - p} \tag{84}$$

where as in Proposition 9.1 $\tilde{Q}_Z$ is the orthogonal complement of Q_Z. Now

$$y \sim N(X\beta, \sigma^2 I_n) \tag{85}$$

If $Q = [Q_Z, \tilde{Q}_Z]$ as in Proposition 9.1 then it follows that:

$$\begin{aligned} Q^T y &= \begin{pmatrix} Q_Z^T y \\ \tilde{Q}_Z^T y \end{pmatrix} \\ &\sim N\left(\begin{pmatrix} Q_Z^T X\beta \\ \tilde{Q}_Z^T X\beta \end{pmatrix}, \sigma^2 \begin{pmatrix} Q_Z^T Q_Z & Q_Z^T \tilde{Q}_Z \\ \tilde{Q}_Z^T Q_Z & \tilde{Q}_Z^T \tilde{Q}_Z \end{pmatrix} \right) \end{aligned} \tag{86}$$

Using the orthogonality of Q_Z, $\tilde{Q}_Z$ and $\tilde{Q}_Z^T \tilde{Q} = \mathbf{0}$ we can write:

$$\begin{aligned} Q^T y &= \begin{pmatrix} Q_Z^T y \\ \tilde{Q}_Z^T y \end{pmatrix} \\ &\sim N\left(\begin{pmatrix} Q_Z^T X\beta \\ \tilde{Q}_Z^T X\beta \end{pmatrix}, \sigma^2 \begin{pmatrix} I_p & \mathbf{0} \\ \mathbf{0} & I_{n-p} \end{pmatrix} \right) \end{aligned} \tag{87}$$

From 87 it is clear that $Q_Z^T y$ and $\tilde{Q}_Z^T y$ have a joint distribution that is multivariate Gaussian with 0 cross-covariance. Hence $Q_Z^T y$ and $\tilde{Q}_Z^T y$ are independent. Inspection of 83 and 84 reveals that $\hat{\gamma}$ is a function only of $Q_Z^T y$ and $\hat{\sigma}_1^2$ is a function only of $\tilde{Q}_Z^T y$. By the independence of $Q_Z^T y$ and $\tilde{Q}_Z^T y$ it is clear that $\hat{\gamma}$ and $\hat{\sigma}_1^2$ are functions of two independent sets of random variables. Therefore $\hat{\gamma}$ and $\hat{\sigma}_1^2$ are independent. ▮■

Proposition 9.3: Suppose $c_X \in \mathbf{R}^q$ is the contrast of interest relative to matrix X and $c_Z \in \mathbf{R}^p$ is the corresponding contrast relative to matrix Z. Suppose β is any vector that satisfies $c_X^T \beta = 0$.

(1) Given β, if there exist vectors $u \in \mathbf{R}^p$ and $v \in \mathbf{R}^n$ such that

$$\begin{aligned} X \left[I_q - c_X (c_X^T c_X)^{-1} c_X^T \right] \beta = {} & Z \left[I_p - c_Z (c_Z^T c_Z)^{-1} c_Z^T \right] u \\ & + \left[I_n - Z (Z^T Z)^{-1} Z^T \right] v \end{aligned} \tag{88}$$

then $E(c_Z^T \hat{\gamma}) = 0$.

(2) Further, if $v = \mathbf{0}$ in equation 88 then the non-centrality parameter $\Delta = 0$ in 68 where P_Z is defined in 70.

Proof: Suppose $c_X^T \beta = 0$ then

$$X\beta = X \left[I_q - c_X (c_X^T c_X)^{-1} c_X^T \right] \beta \tag{89}$$

Replacing the right hand side of 89 by the right hand side of 88 we see that when $c_X^T \beta = 0$:

$$X\beta = Z \left[I_p - c_Z (c_Z^T c_Z)^{-1} c_Z^T \right] u + \left[I_n - Z (Z^T Z)^{-1} Z^T \right] v \tag{90}$$

First, we prove (1). Left multiplying both sides by $c_Z^T (Z^T Z)^{-1} Z^T$ and simplifying we get:

$$\begin{aligned} c_Z^T (Z^T Z)^{-1} Z^T X\beta &= \left[c_Z^T - c_Z^T \right] u + c_Z^T (Z^T Z)^{-1} \left[Z^T - Z^T \right] v \\ &= 0 \end{aligned} \tag{91}$$

From Proposition 9.1 $E(\hat{\gamma}) = (Z^T Z)^{-1} Z^T X\beta$ and so $E(c_Z^T \hat{\gamma}) = c_Z^T (Z^T Z)^{-1} Z^T X\beta$. Therefore it follows from 91 that $E(c_Z^T \hat{\gamma}) = 0$ which proves (1).

Next, we prove (2). Left multiplying both sides of 90 by P_Z we get:

$$\begin{aligned} P_Z X\beta = {} & P_Z \, Z \left[I_p - c_Z (c_Z^T c_Z)^{-1} c_Z^T \right] u \\ & + P_Z \, \left[I_n - Z (Z^T Z)^{-1} Z^T \right] v \end{aligned} \tag{92}$$

Note that $P_Z \, Z = [I_n - Z(Z^T Z)^{-1} Z^T] Z = [Z - Z] = \mathbf{0}$ and $v = \mathbf{0}$ as per the given condition in (2). Thus $P_Z X\beta = \mathbf{0}$. From the definition of Δ it follows that $\Delta = \frac{\beta^T X^T \, (P_Z \, X\beta)}{\sigma^2} = 0$ which proves (2).

■

Corollary 9.4: Suppose $X = [x_1, X_2]$ where $x_1 \in \mathbf{R}^n$, $X_2 \in \mathbf{R}^{n \times (q-1)}$ and $c_X = [1, 0, \ldots, 0]^T \in \mathbf{R}^q$. Let Z be any matrix such that $Z = [z_1, X_2, Z_3]$ where $z_1 \in \mathbf{R}^n$, $Z_3 \in \mathbf{R}^{n \times (p-q)}$ and let $c_Z = [1, 0, \ldots, 0]^T \in \mathbf{R}^p$. Then for any β such that $c_X^T \beta = 0$ we have $E(c_Z^T \hat{\gamma}) = 0$ and $\Delta = 0$.

Proof: First note that for $c_X = [1, 0, \ldots, 0]^T$

$$X \left[I_q - c_X (c_X^T c_X)^{-1} c_X^T \right] = [0, X_2] \tag{93}$$

and similarly for $c_Z = [1, 0, \ldots, 0]^T$

$$Z \left[I_p - c_Z (c_Z^T c_Z)^{-1} c_Z^T \right] = [0, X_2, Z_3] \tag{94}$$

Let us partition $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \in \mathbf{R}^q$ where $\beta_1 \in \mathbf{R}$ and $\beta_2 \in \mathbf{R}^{(q-1)}$. Now the vector $\begin{pmatrix} 0 \\ \beta_2 \end{pmatrix}$ satisfies $c_X^T \beta = 0$ for any choice of β_2. Choose $v = \mathbf{0} \in \mathbf{R}^n$ and $u = \begin{pmatrix} 0 \\ \beta_2 \\ g \end{pmatrix}$ where $g = \mathbf{0} \in \mathbf{R}^{(p-q)}$. With this choice the left hand side of 88 is

$$X \left[I_q - c_X (c_X^T c_X)^{-1} c_X^T \right] \beta = [0, X_2] \begin{pmatrix} 0 \\ \beta_2 \end{pmatrix} = X_2 \beta_2 \tag{95}$$

and the right hand side of 88 is

$$\begin{aligned} & Z \left[I_p - c_Z (c_Z^T c_Z)^{-1} c_Z^T \right] u + \left[I_n - Z (Z^T Z)^{-1} Z^T \right] v \\ & \quad = [0, X_2, Z_3] \begin{pmatrix} 0 \\ \beta_2 \\ g \end{pmatrix} = X_2 \beta_2 \end{aligned} \tag{96}$$

Thus 88 can always be satisfied by choosing u and v appropriately with $v = \mathbf{0} \in \mathbf{R}^n$ for any given β satisfying $c_X^T \beta = 0$ and hence by Proposition 9.3 $E(c_Z^T \hat{\gamma}) = 0$ and $\Delta = 0$. ▮

■

C. Derivation of gradient equations

Consider the matrix $A \in \mathbf{R}^{n\times m}$, square matrices $B, C \in \mathbf{R}^{n\times n}$ and vectors $x \in \mathbf{R}^n$ and $y, z \in \mathbf{R}^m$. We will use tr to denote the trace operator on a square matrix. Recall that the matrix derivative with respect to matrix A of a scalar quantity f is denoted by $\frac{\partial f}{\partial A}$. The ijth element of $\frac{\partial f}{\partial A}$ is defined as:

$$\left[\frac{\partial f}{\partial A}\right]_{ij} = \frac{\partial f}{\partial a_{ij}} \tag{97}$$

where a_{ij} is the ijth element of A.

Below we note some basic identities:

$$\left[\frac{\partial}{\partial A}(x^TAy)\right]_{ij} = \left[\frac{\partial}{\partial A}(y^TA^Tx)\right]_{ij} = \left[xy^T\right]_{ij} \tag{98}$$

$$\left[\frac{\partial}{\partial B}(x^TB^{-1}x)\right]_{ij} = \left[-B^{-T}xx^TB^{-T}\right]_{ij} \tag{99}$$

$$\begin{aligned}&\left[\frac{\partial}{\partial A}(y^T(A^TA)^{-1}z)\right]_{ij}\\ &= \left[-A(A^TA)^{-1}(zy^T+yz^T)(A^TA)^{-1}\right]_{ij}\end{aligned} \tag{100}$$

$$\left[\frac{\partial}{\partial B}\text{tr}(BC)\right]_{ij} = \left[C^T\right]_{ij} \tag{101}$$

$$\left[\frac{\partial}{\partial B}\text{tr}(B^TC)\right]_{ij} = \left[C\right]_{ij} \tag{102}$$

$$\text{tr }(BC) = \text{ tr }(CB) \tag{103}$$

Recall, equation 29, the objective function under consideration:

$$\begin{aligned}&G_\phi(Z, c_Z)\\ &\quad= c_Z^T(Z^TZ)^{-1}c_Z\left[\sum_{i=1}^m 2\phi_i w_i + tr\left(P_ZH\Phi_V\Sigma H^T\right)\right]\\ &\quad+ c_Z^T(Z^TZ)^{-1}Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Z\\ &\quad- 2c_Z^T(Z^TZ)^{-1}Z^TH\Phi_B\ell\\ &\quad+\sum_{i=1}^m w_i(2-2\phi_i)(c_{X_i}^T\beta_i/\sigma_i)^2\end{aligned} \tag{104}$$

The matrix gradient with respect to Z can be written as:

$$\begin{aligned}&\frac{\partial}{\partial Z}G_\phi(Z, c_Z)\\ &= \frac{\partial}{\partial Z}\left\{c_Z^T(Z^TZ)^{-1}c_Z\right\}\left[\sum_{i=1}^m 2\phi_i w_i + tr\left(P_ZH\Phi_V\Sigma H^T\right)\right]\\ &+ c_Z^T(Z^TZ)^{-1}c_Z\frac{\partial}{\partial Z}\left\{\left[\sum_{i=1}^m 2\phi_i w_i + tr\left(P_ZH\Phi_V\Sigma H^T\right)\right]\right\}\\ &+\frac{\partial}{\partial Z}\left\{c_Z^T(Z^TZ)^{-1}Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Z\right\}\\ &+\frac{\partial}{\partial Z}\left\{-2c_Z^T(Z^TZ)^{-1}Z^TH\Phi_B\ell\right\}\\ &+\frac{\partial}{\partial Z}\left\{\sum_{i=1}^m w_i(2-2\phi_i)(c_{X_i}^T\beta_i/\sigma_i)^2\right\}\end{aligned} \tag{105}$$

For the sake of presentation clarity, we define the following smaller terms in equation 135:

$$\begin{aligned}&\textbf{Term 1} = \frac{\partial}{\partial Z}\left\{c_Z^T(Z^TZ)^{-1}c_Z\right\}\\ &\textbf{Term 2} = \frac{\partial}{\partial Z}\left\{\left[tr\left(P_ZH\Phi_V\Sigma H^T\right)\right]\right\}\\ &\textbf{Term 3} = \frac{\partial}{\partial Z}\left\{c_Z^T(Z^TZ)^{-1}Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Z\right\}\\ &\textbf{Term 4} = \frac{\partial}{\partial Z}\left\{-2c_Z^T(Z^TZ)^{-1}Z^TH\Phi_B\ell\right\}\\ &\textbf{Term 5} = \frac{\partial}{\partial Z}\left\{\sum_{i=1}^m w_i(2-2\phi_i)(c_{X_i}^T\beta_i/\sigma_i)^2\right\}\end{aligned} \tag{106}$$

*1) **Term 1**:* From 100 we have:

$$\frac{\partial}{\partial Z}\left\{c_Z^T(Z^TZ)^{-1}c_Z\right\} = -Z(Z^TZ)^{-1}(2c_Zc_Z^T)(Z^TZ)^{-1} \tag{107}$$

*2) **Term 2**:*

$$\begin{aligned}&\frac{\partial}{\partial Z_{ij}}\left\{\left[tr\left(P_ZH\Phi_V\Sigma H^T\right)\right]\right\}\\ &\quad= tr\left(-\frac{\partial Z}{\partial Z_{ij}}(Z^TZ)^{-1}Z^TH\Phi_V\Sigma H^T\right)\\ &\quad+ tr\left(-Z\frac{\partial}{\partial Z_{ij}}\left\{(Z^TZ)^{-1}\right\}Z^TH\Phi_V\Sigma H^T\right)\\ &\quad+ tr\left(-Z(Z^TZ)^{-1}\frac{\partial Z^T}{\partial Z_{ij}}H\Phi_V\Sigma H^T\right)\end{aligned} \tag{108}$$

Again,

$$\begin{aligned}&tr\left(-\frac{\partial Z}{\partial Z_{ij}}(Z^TZ)^{-1}Z^TH\Phi_V\Sigma H^T\right)\\ &\quad= -\left[H\Sigma^T\Phi_V^TH^TZ(Z^TZ)^{-1}\right]_{ij}\end{aligned} \tag{109}$$

and

$$\begin{aligned}&tr\left(-Z(Z^TZ)^{-1}\frac{\partial Z^T}{\partial Z_{ij}}H\Phi_V\Sigma H^T\right)\\ &\quad= tr\left(-\frac{\partial Z^T}{\partial Z_{ij}}H\Phi_V\Sigma H^TZ(Z^TZ)^{-1}\right)\\ &\quad= -\left[H\Phi_V\Sigma H^TZ(Z^TZ)^{-1}\right]_{ij}\end{aligned} \tag{110}$$

Also,

$$\begin{aligned}&tr\left(-Z\frac{\partial}{\partial Z_{ij}}\left\{(Z^TZ)^{-1}\right\}Z^TH\Phi_V\Sigma H^T\right)\\ &= tr\left(Z(Z^TZ)^{-1}\frac{\partial Z^T}{\partial Z_{ij}}Z(Z^TZ)^{-1}Z^TH\Phi_V\Sigma H^T\right)\\ &+ tr\left(Z(Z^TZ)^{-1}Z^T\frac{\partial Z}{\partial Z_{ij}}(Z^TZ)^{-1}Z^TH\Phi_V\Sigma H^T\right)\\ &= tr\left(\frac{\partial Z^T}{\partial Z_{ij}}Z(Z^TZ)^{-1}Z^TH\Phi_V\Sigma H^TZ(Z^TZ)^{-1}\right)\\ &+ tr\left(\frac{\partial Z}{\partial Z_{ij}}(Z^TZ)^{-1}Z^TH\Phi_V\Sigma H^TZ(Z^TZ)^{-1}Z^T\right)\\ &= \left[Z(Z^TZ)^{-1}Z^TH\Phi_V\Sigma H^TZ(Z^TZ)^{-1}\right]_{ij}\\ &+ \left[Z(Z^TZ)^{-1}Z^TH\Sigma^T\Phi_V^TH^TZ(Z^TZ)^{-1}\right]_{ij}\end{aligned} \tag{111}$$

Combining 139, 140, 110 and 111 we get:

$$\begin{aligned}
&\frac{\partial}{\partial Z}\left\{\left[tr\left(P_Z H\Phi_V\Sigma H^T\right)\right]\right\} \\
&= -H\Sigma^T\Phi_V^T H^T Z(Z^TZ)^{-1} \\
&+ Z(Z^TZ)^{-1}Z^T H\Phi_V\Sigma H^T Z(Z^TZ)^{-1} \\
&+ Z(Z^TZ)^{-1}Z^T H\Sigma^T\Phi_V^T H^T Z(Z^TZ)^{-1} \\
&- H\Phi_V\Sigma H^T Z(Z^TZ)^{-1} \\
&= -(H\Phi_V\Sigma H^T + H\Sigma^T\Phi_V^T H^T)Z(Z^TZ)^{-1} \\
&+ Z(Z^TZ)^{-1}Z^T(H\Phi_V\Sigma H^T + H\Sigma^T\Phi_V^T H^T)Z(Z^TZ)^{-1} \\
&= -P_Z(H\Phi_V\Sigma H^T + H\Sigma^T\Phi_V^T H^T)Z(Z^TZ)^{-1}
\end{aligned} \tag{112}$$

*3) **Term 3**:*

$$\begin{aligned}
&\frac{\partial}{\partial Z_{ij}}\left\{c_Z^T(Z^TZ)^{-1}Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Z\right\} \\
&= c_Z^T\frac{\partial}{\partial Z_{ij}}\left\{(Z^TZ)^{-1}\right\}Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Z \\
&+c_Z^T(Z^TZ)^{-1}\frac{\partial}{\partial Z_{ij}}\left\{Z^T\right\}H\Phi_BH^TZ(Z^TZ)^{-1}c_Z \\
&+c_Z^T(Z^TZ)^{-1}Z^TH\Phi_BH^T\frac{\partial}{\partial Z_{ij}}\left\{Z\right\}(Z^TZ)^{-1}c_Z \\
&+c_Z^T(Z^TZ)^{-1}Z^TH\Phi_BH^TZ\frac{\partial}{\partial Z_{ij}}\left\{(Z^TZ)^{-1}\right\}c_Z
\end{aligned} \tag{113}$$

Applying 126 and 100 repeatedly to each of the above terms we get:

$$\begin{aligned}
&\frac{\partial}{\partial Z}\left\{c_Z^T(Z^TZ)^{-1}Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Z\right\} \\
&= -Z(Z^TZ)^{-1}(Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Zc_Z^T)(Z^TZ)^{-1} \\
&- Z(Z^TZ)^{-1}(c_Zc_Z^T(Z^TZ)^{-1}Z^TH\Phi_B^TH^TZ)(Z^TZ)^{-1} \\
&+ H\Phi_BH^TZ(Z^TZ)^{-1}c_Zc_Z^T(Z^TZ)^{-1} \\
&+ H\Phi_B^TH^TZ(Z^TZ)^{-1}c_Zc_Z^T(Z^TZ)^{-1} \\
&- Z(Z^TZ)^{-1}(c_Zc_Z^T(Z^TZ)^{-1}Z^TH\Phi_BH^TZ)(Z^TZ)^{-1} \\
&- Z(Z^TZ)^{-1}(Z^TH\Phi_B^TH^TZ(Z^TZ)^{-1}c_Zc_Z^T)(Z^TZ)^{-1}
\end{aligned} \tag{114}$$

Since Φ_B is diagonal, this can be simplified to:

$$\begin{aligned}
&\frac{\partial}{\partial Z}\left\{c_Z^T(Z^TZ)^{-1}Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Z\right\} \\
&= -2Z(Z^TZ)^{-1}(Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Zc_Z^T)(Z^TZ)^{-1} \\
&- 2Z(Z^TZ)^{-1}(c_Zc_Z^T(Z^TZ)^{-1}Z^TH\Phi_B^TH^TZ)(Z^TZ)^{-1} \\
&+ 2H\Phi_BH^TZ(Z^TZ)^{-1}c_Zc_Z^T(Z^TZ)^{-1}
\end{aligned} \tag{115}$$

*4) **Term 4**:*

$$\begin{aligned}
&\frac{\partial}{\partial Z_{ij}}\left\{-2c_Z^T(Z^TZ)^{-1}Z^TH\Phi_B\ell\right\} \\
&\quad= \left\{-2c_Z^T\frac{\partial}{\partial Z_{ij}}\left\{(Z^TZ)^{-1}\right\}Z^TH\Phi_B\ell\right\} \\
&\quad+ \left\{-2c_Z^T(Z^TZ)^{-1}\frac{\partial}{\partial Z_{ij}}\left\{Z^T\right\}H\Phi_B\ell\right\}
\end{aligned} \tag{116}$$

Application of 126 and 100 gives:

$$\begin{aligned}
&\frac{\partial}{\partial Z}\left\{-2c_Z^T(Z^TZ)^{-1}Z^TH\Phi_B\ell\right\} \\
&= (-2)(-Z)(Z^TZ)^{-1}(Z^TH\Phi_B\ell c_Z^T + c_Z\ell^T\Phi_B^TH^TZ)(Z^TZ)^{-1} \\
&+ (-2)H\Phi_B\ell c_Z^T(Z^TZ)^{-1}
\end{aligned} \tag{117}$$

*5) **Term 5**:* Term 5 is a constant w.r.t Z and c_Z and so the gradient w.r.t Z and c_Z is 0.

6) Combining terms: Combining 135, 138, 112, 115 and 117 gives:

$$\begin{aligned}
&\frac{\partial}{\partial Z}G_\phi(Z,c_Z) \\
&= \left[-Z(Z^TZ)^{-1}(2c_Zc_Z^T)(Z^TZ)^{-1}\right]\left[\sum_{i=1}^{m}2\phi_iw_i\right] \\
&+ \left[-Z(Z^TZ)^{-1}(2c_Zc_Z^T)(Z^TZ)^{-1}\right]\left[tr\left(P_ZH\Phi_V\Sigma H^T\right)\right] \\
&+ c_Z^T(Z^TZ)^{-1}c_Z\left[-P_Z(H\Phi_V\Sigma H^T)Z(Z^TZ)^{-1}\right] \\
&+ c_Z^T(Z^TZ)^{-1}c_Z\left[-P_Z(H\Sigma^T\Phi_V^TH^T)Z(Z^TZ)^{-1}\right] \\
&+ \left[-2Z(Z^TZ)^{-1}(Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Zc_Z^T)(Z^TZ)^{-1}\right] \\
&+ \left[-2Z(Z^TZ)^{-1}(c_Zc_Z^T(Z^TZ)^{-1}Z^TH\Phi_B^TH^TZ)(Z^TZ)^{-1}\right] \\
&+ \left[2H\Phi_BH^TZ(Z^TZ)^{-1}c_Zc_Z^T(Z^TZ)^{-1}\right] \\
&+ \left[(-2)(-Z)(Z^TZ)^{-1}(Z^TH\Phi_B\ell c_Z^T + c_Z\ell^T\Phi_B^TH^TZ)(Z^TZ)^{-1}\right] \\
&+ \left[(-2)H\Phi_B\ell c_Z^T(Z^TZ)^{-1}\right]
\end{aligned} \tag{118}$$

Noting that Φ_B, Φ_V and Σ are diagonal and given the fact that diagonal matrices commute, rearrangement gives:

$$\begin{aligned}
&\frac{\partial}{\partial Z}G_\phi(Z,c_Z) \\
&= \left[-Z(Z^TZ)^{-1}(2c_Zc_Z^T)(Z^TZ)^{-1}\right]\left[\sum_{i=1}^{m}2\phi_iw_i\right] \\
&+ \left[-Z(Z^TZ)^{-1}(2c_Zc_Z^T)(Z^TZ)^{-1}\right]\left[tr\left(P_ZH\Phi_V\Sigma H^T\right)\right] \\
&- 2\left(c_Z^T(Z^TZ)^{-1}c_Z\right)\left[P_Z(H\Phi_V\Sigma H^T)Z(Z^TZ)^{-1}\right] \\
&- 2Z(Z^TZ)^{-1}\left[c_Zc_Z^T(Z^TZ)^{-1}Z^TH\Phi_BH^TZ\right](Z^TZ)^{-1} \\
&- 2Z(Z^TZ)^{-1}\left[Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Zc_Z^T\right](Z^TZ)^{-1} \\
&+ \left[2H\Phi_BH^TZ(Z^TZ)^{-1}c_Zc_Z^T(Z^TZ)^{-1}\right] \\
&+ 2Z(Z^TZ)^{-1}(Z^TH\Phi_B\ell c_Z^T)(Z^TZ)^{-1} \\
&+ 2Z(Z^TZ)^{-1}(c_Z\ell^T\Phi_B^TH^TZ)(Z^TZ)^{-1} \\
&- 2H\Phi_B\ell c_Z^T(Z^TZ)^{-1}
\end{aligned} \tag{119}$$

Thus the gradient with respect to c_Z is easily computed as:

$$\begin{aligned}
&\frac{\partial}{\partial c_Z}G_\phi(Z,c_Z) \\
&\quad= 2(Z^TZ)^{-1}c_Z\left[\sum_{i=1}^{m}2\phi_iw_i + tr\left(P_ZH\Phi_V\Sigma H^T\right)\right] \\
&\quad+ 2(Z^TZ)^{-1}Z^TH\Phi_BH^TZ(Z^TZ)^{-1}c_Z \\
&\quad- 2(Z^TZ)^{-1}Z^TH\Phi_B\ell
\end{aligned} \tag{120}$$

Equations 119 and 120 are the same as 40 and 41 respectively.

D. Additional examples

The reader is referred to Example 1 IV-A in the main text for details about the parameters of the following 2 case studies.

1) Example 2: This example is the same as Example 1 from the main text except for two differences:

1) First we choose ϕ_i automatically using the strategy proposed in section III-D. We used Option A to initialize ϕ_i using $k = 10^3$.
2) Second, we initialize the columns of Z automatically using the strategy proposed in section III-B.

The results are shown in Figures 27 and 28.

2) Example 3: This example is similar to Example 1 from the main text. Here we attempt to investigate the effect of bias-variance weightings on the performance curves. We put a higher emphasis on reducing bias by choosing $\phi = 0.1$. Please note that the relationship between the value of ϕ_i and the importance of bias or variance terms is non-linear (see section III-D). The first two columns of Z were fixed as before and the contrast c_Z was fixed at [1;0;0]. We also chose $w_i = 1$ to give equal weights to all design matrices. The unconstrained column in Z was initialized randomly with elements drawn from a uniform distribution $U(0, 1)$. The results are shown in Figures 29 and 30.

E. Validation

We validate the proposed practical algorithm by comparing optimal solutions from the more sophisticated solver with the ones produced using the practical solver Algorithm 1. Our state-of-the-art optimization solver (see appendix IX-F) was used to solve the validation test problems.The optimization core uses an augmented lagrangian algorithm (inspired by the implementation in LANCELOT package [16], [17]) to solve equality constrained problems. Inequality constraints are handled by first transforming them to equality constraints via slack variables and solving the resulting bound constrained optimization problem. Some features of interest are as follows:

1) A Trust region based approach [18] is used to generate search directions at each step (for both equality constrained and inequality constrained problems).
2) For equality constraints only, the subproblems above are solved using a conjugate gradient approach (Newton-CG -Steihaug) [19] that is fast and accurate even for large problems and can handle both positive definite and indefinite Hessian approximations. If both equality and inequality constraints are present then we solve the trust region problem with a non-linear gradient projection technique [20] followed by subspace optimization using Newton-CG-Steihaug.
3) A symmetric rank 1 (SR1) quasi-Newton approximation to the Hessian [21] is used which is known to generate good Hessian approximations for both convex and non-convex problems. As suggested in [22] we do the update also on the rejected steps to gather curvature information about the function. We provide options for BFGS [23] especially for convex problems and an option for preconditioning the CG iterations. We also implement limited memory variants of SR1 and BFGS for large problems.
4) Our algorithm accepts vectorized constraints so that multiple constraints can be programmed simultaneously. Only gradient information is required. Hessian information is optional but not required.

We tested the performance of our algorithm using standard optimization benchmarks from the GAMS performance benchmark problems (http://www.gamsworld.org/performance, [24]). The appendix IX-F provides more technical details of the algorithm.

1) Validation Test A: Motivated by a practical data-set that we later describe we consider for illustration purposes the case study when the profiles of interest are shifted relative to a base profile by variable units and our goal is to simultaneously capture all responses with a single design matrix. To test and validate the optimization framework, we used the basic design matrix X_0 from Figure 20. Fifty ($m = 50$) expected design matrices were proposed with:

$$X_i(:,1) = X_0(:,1) \text{ shifted right by } i \text{ timepoints} \tag{121}$$

$$X_i(:,2) = X_0(:,2) \tag{122}$$

We chose $\frac{\beta_i}{\sigma_i} = [1, 0.5]^T$ and $c_{X_i} = [1, 0]^T, \forall i$. The weights were chosen as $w_i = 1, \forall i$ to reflect the equal likelihood of observing any X_i. We chose $\phi_i = 0.5, \forall i$ in this validation test.

The rank of Z was chosen to be 4 and the matrix A was chosen as $A = [e_1, e_2]$ where $e_1 \in R^4$ is a unit vector with 1 at position 1 and zeros elsewhere. Similarly for e_2. The matrix B was chosen as X_0 to fix the first two columns of Z to those of X_0. C was chosen as the identity matrix I_4 and d was set to $[1, 0, 0, 0]^T$ to fix the contrast c_Z.

2) Validation Test B: In this case, the contrast vector c_Z was left unconstrained. Everything else is the same as in Validation Test A. Convergence diagnostics and optimal $\hat{Z}$ for this case are shown in Figure 33 and Figure 34 respectively. The optimal contrast was determined to be $\hat{c_Z} = [0.73519; 0.47890; 0.76016; 0.75789]$.

TABLE II
OPTIMAL OBJECTIVE VALUES FOR THE EXACT ALGORITHM AND ALGORITHM 1 FOR CASE A AND CASE B.

	Exact	Algorithm 1
Case a	2.693467	2.693490
Case b	0.265484	0.265502

The optimal contrast using Algorithm 1 was found to be $\hat{c_Z} = [0.73523; 0.47890; 0.75704; 0.75844]$.

F. Exact solver details

Our optimization algorithm solves the general problem:

$$\min\ _x f(x) \tag{123}$$

$$\text{s.t. } c_i(x) = 0, \qquad i = 1, 2, \ldots, m \tag{124}$$

$$\text{s.t. } g_j(x) \geq 0, \qquad j = 1, 2, \ldots L \tag{125}$$

where $x \in R^n$.

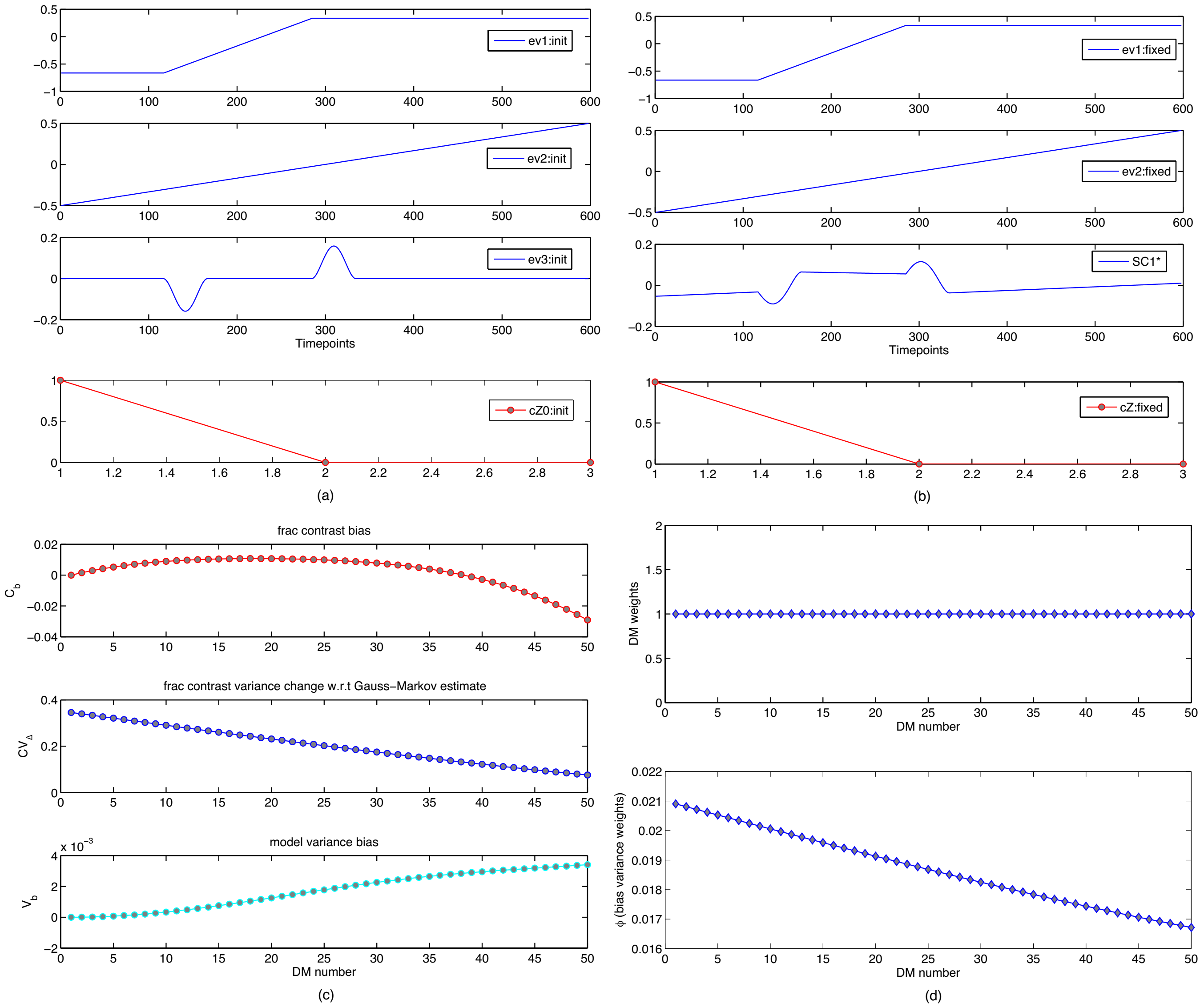


Fig. 27. Example 2: (a) Initial design matrix (DM) initialized using the automatic strategy proposed in section III-B. The first two columns were fixed at their initial values and the contrast was fixed at [1;0;0] (b) Estimated optimal DM. (c) Performance curves showing the fractional contrast bias C_b, contrast variance change w.r.t Gauss-Markov estimate $CV_\triangle$ and model variance bias V_b (d) For all i, the DM weights $w_i = 1$ implying equal likelihood of observing any of the specified DM's. The bias-variance scalings ϕ_i were chosen using the automatic strategy proposed in section III-D using $k = 10^3$.

We convert the inequality constraints into equality constraints via slack variables as follows:

$$g_j(x) - s_j = 0 \tag{126}$$
$$s_j \geq 0, \quad j = 1, 2, \dots L \tag{127}$$

Thus the optimization problem becomes:

$$\min \ f(x) \tag{128}$$
$$\text{s.t. } c_i(x) = 0, \qquad i = 1, 2, \dots, m \tag{129}$$
$$\text{s.t. } g_j(x) - s_j = 0, \qquad j = 1, 2, \dots L \tag{130}$$
$$s_j \geq 0 \tag{131}$$

This problem is now an equality constrained problem where the inequalities have been replaced by the bound constraints on the slack variables. Thus it suffices to consider equality constrained problems with bounds on independent variables as follows:

$$\min \ f(x) \tag{132}$$
$$\text{s.t. } c_i(x) = 0, \qquad i = 1, 2, \dots, m \tag{133}$$
$$\text{s.t. } l_i \leq x_i \leq u_i, \qquad i = 1, 2, \dots n \tag{134}$$

where $x \in R^n$.

Our code uses a trust region based augmented lagrangian approach to solve these bound constrained problems following closely the LANCELOT software package [17], [16]. The augmented lagrangian function for the above problem is defined as:

$$\mathcal{L}(x, \lambda, \mu) = f(x) - \sum_{i=1}^{m} \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i=1}^{m} c_i(x)^2 \tag{135}$$

At each outer iteration k, given current values of λ^k and μ_k we solve the subproblem:

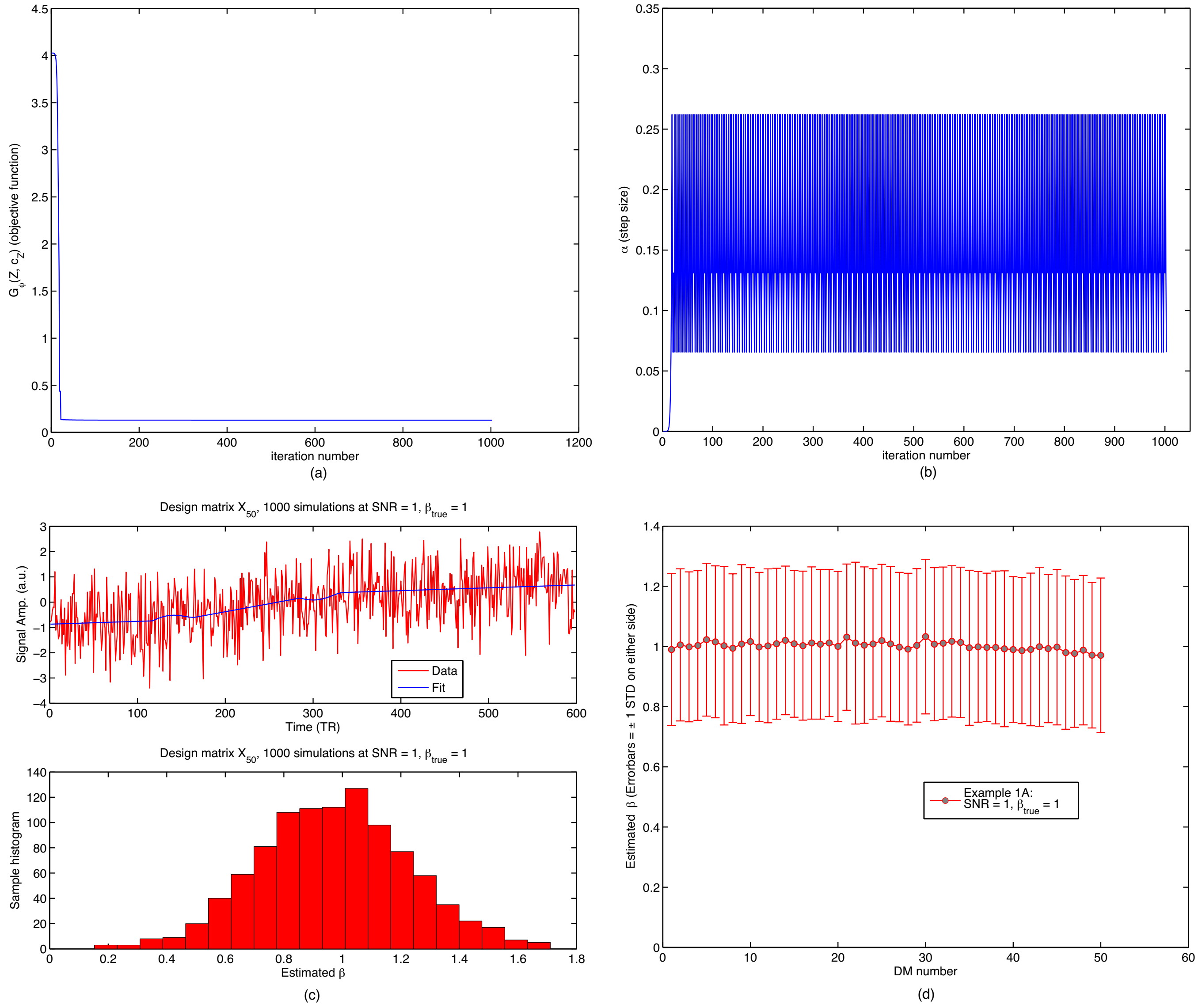


Fig. 28. Example 2: (a) Figure showing the evolution of objective function values $G_\phi(Z, c_Z)$ over algorithm iterations. Notice how the function value stabilizes as convergence is reached (b) Figure showing the variation in the step size α over algorithm iterations. Step size controlling parameter θ in Algorithm 1 was set to $\theta = 2$. For each design matrix (DM) X_i entered into optimization, 1000 simulated data-sets were generated at SNR $\frac{\beta_i}{\sigma_i}$. A GLM analysis was run on each of these data-sets using the optimized DM. Figure (c) shows an example of simulated data for DM X_{50} at SNR $\frac{\beta_{50}}{\sigma_{50}}$ and the GLM fit using the optimal DM. It also shows the distribution of $c_Z^T\hat{\gamma}$ over 1000 simulations. Figure (d) is a summary errorbar plot showing $\hat{E}(c_Z^T\hat{\gamma})$ over 1000 simulations for data generated from each DM. The error bars represent unit standard deviation of $c_Z^T\hat{\gamma}$ (**not** standard deviation of $\hat{E}(c_Z^T\hat{\gamma})$) to quantify the variance in estimation via simulation.

$$\min \mathcal{L}(x, \lambda^k, \mu_k) \tag{136}$$

$$\text{s.t. } l_i \leq x_i \leq u_i \tag{137}$$

If P is the projection operator defined as

$$[P(z, l, u)]_i = \begin{cases} l_i & \text{if} \quad z_i \leq l_i \\ z_i & \text{if} \quad l_i \leq z_i \leq u_i \\ u_i & \text{if} \quad z_i \geq u_i \end{cases} \tag{138}$$

then the Karush-Kuhn-Tucker (KKT) optimality condition for 136 is given as [16]:

$$x - P(x - \nabla_x \mathcal{L}(x, \lambda^k, \mu_k), l, u) = 0 \tag{139}$$

The outer iteration code is given in Framework 1. Note that the penalty parameter μ_k is updated based on a feasibility monitoring strategy that allows for a decrease in μ_k if sufficient accuracy is not achieved in solving the subproblem 136.

At each inner iteration we form a quadratic approximation to the augmented lagrangian and approximately solve the inequality constrained quadratic sub-problem:

$$\min{}_p \ \frac{1}{2} p^T \nabla^2_{xx} \mathcal{L}(x, \lambda, \mu) p + \nabla_x \mathcal{L}(x, \lambda, \mu)^T p \tag{140}$$

$$\text{s.t. } l_i \leq x_i \leq u_i \tag{141}$$

$$\text{s.t. } ||p||_\infty \leq \Delta \tag{142}$$

The inner iteration code uses non-linear gradient projection

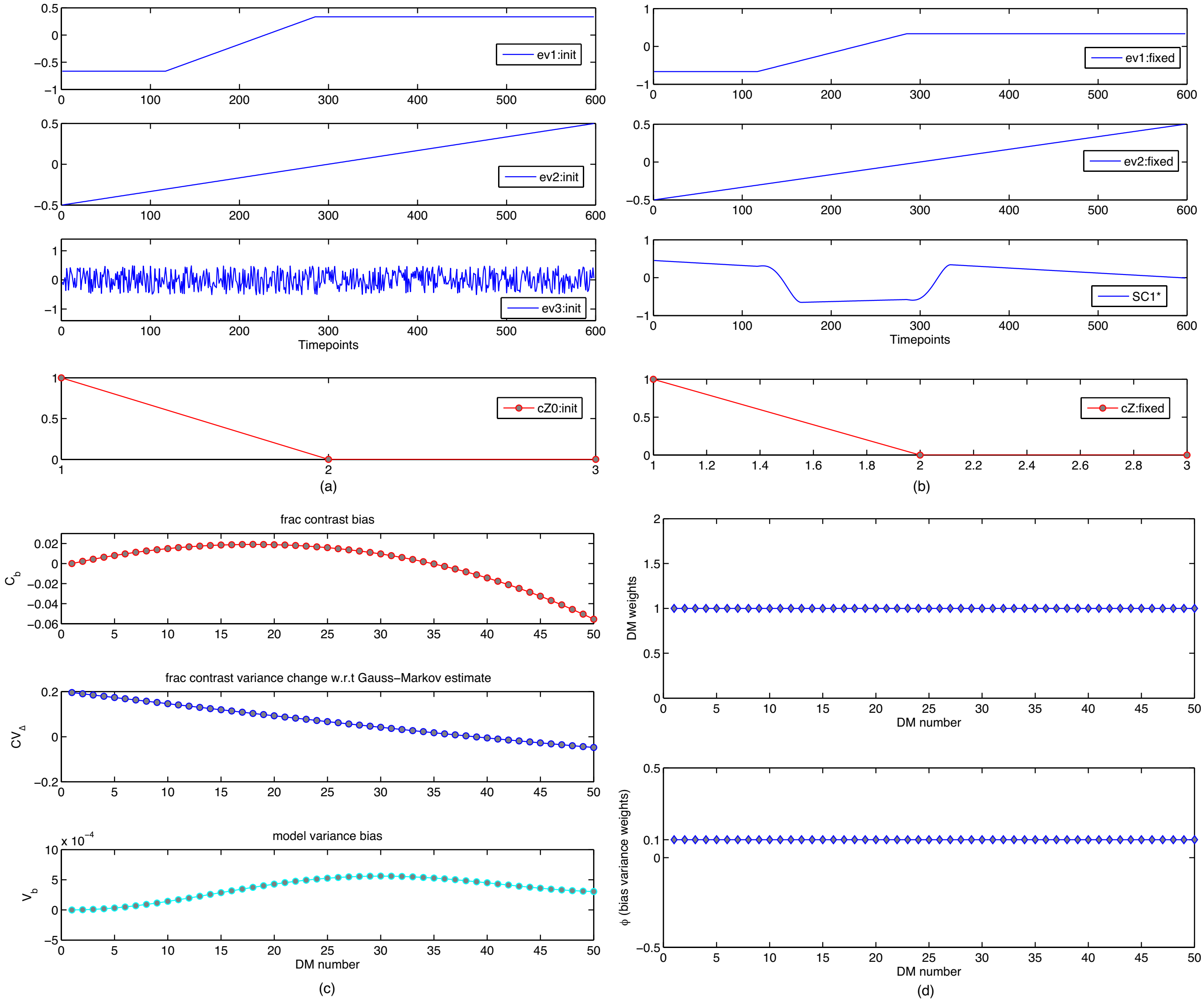


Fig. 29. Example 3: (a) Initial design matrix (DM) along with random initialization of the 3rd column. The first two columns were fixed at their initial values and the contrast was fixed at [1;0;0] (b) Estimated optimal DM. Notice how the 3rd column converges to a non-random profile (c) Performance curves showing the fractional contrast bias C_b, contrast variance change w.r.t Gauss-Markov estimate CV_Δ and model variance bias V_b (d) In this example $w_i = 1$ and ϕ_i = 0.1 indicating a higher weighting to bias term during optimization.

[20] followed by Newton-CG-Steihaug conjugate gradient iterations [19]. Quasi-Newton updates are performed using either SR1 [21] (recommended for non-convex functions) or BFGS [23] (recommended for convex functions). For very large problems, we switch to the limited memory variants [25] of these quasi-Newton approximations. The algorithm details are given in Framework 2. The trust region update code is based on a standard progress monitoring strategy [22] and is given in Framework 3.

G. Half-cosine parameterization of HRF

Here we describe the equations used to generate plausible HRF shapes via a 5-parameter half-cosine parameterization. This parameterization is defined as follows:

$$HRF(t) = \begin{cases} F_1(t) & \text{if} \quad 0 \le t \le h_1 \\ F_2(t) & \text{if} \quad h_1 < t \le \sum_{i=1}^{2} h_i \\ F_3(t) & \text{if} \quad \sum_{i=1}^{2} h_i < t \le \sum_{i=1}^{3} h_i \\ F_4(t) & \text{if} \quad \sum_{i=1}^{3} h_i < t \le \sum_{i=1}^{4} h_i \\ F_5(t) & \text{if} \quad t > \sum_{i=1}^{4} h_i \end{cases} \tag{149}$$

where

$$F_1(t) = 0 \tag{150}$$

$$F_2(t) = \cos\left[\frac{\pi}{2} - \frac{\pi}{2h_2}(t - h_1)\right] \tag{151}$$

$$F_3(t) = \cos\left[\left(\frac{\pi}{2h_3} + \frac{\sin^{-1}(f_2)}{h_3}\right)(h_1 + h_2 - t)\right] \tag{152}$$

$$F_4(t) = f\cos\left[\pi - \frac{\pi}{2h_4}(t - h_1 - h_2 - h_3)\right] \tag{153}$$

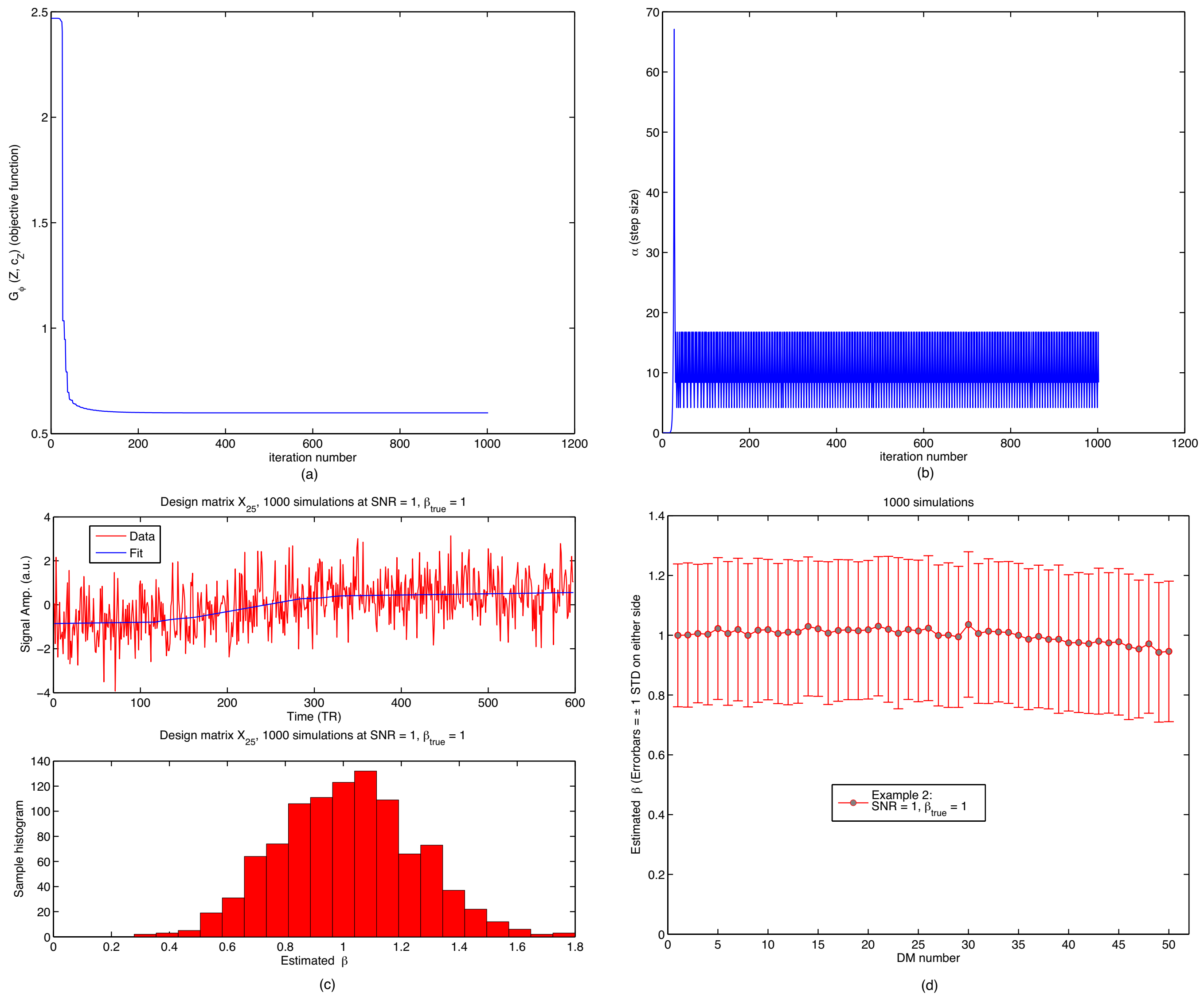


Fig. 30. Example 3: (a) Figure showing the evolution of objective function values $G_\phi(Z, c_Z)$ over algorithm iterations. Notice how the function value stabilizes as convergence is reached (b) Figure showing the variation in the step size α over algorithm iterations. Step size controlling parameter θ in Algorithm 1 was set to $\theta = 2$. For each design matrix (DM) X_i entered into optimization, 1000 simulated data-sets were generated at SNR $\frac{\beta_i}{\sigma_i}$. A GLM analysis was run on each of these data-sets using the optimized DM. Figure (c) shows an example of simulated data for DM X_{25} at SNR $\frac{\beta_{25}}{\sigma_{25}}$ and the GLM fit using the optimal DM. It also shows the distribution of $c_Z^T\hat{\gamma}$ over 1000 simulations. Figure (d) is a summary errorbar plot showing $\hat{E}(c_Z^T\hat{\gamma})$ over 1000 simulations for data generated from each DM. The error bars represent unit standard deviation of $c_Z^T\hat{\gamma}$ (**not** standard deviation of $\hat{E}(c_Z^T\hat{\gamma})$) to quantify the variance in estimation via simulation.

$$F_5(t) = 0 \qquad (154)$$

Acknowledgment

The authors acknowledge financial support from the Imaging consortium for drug development (ICD).

References

[1] K. Friston, A. Holmes, K. Worsley, J.-B. Poline, C. Frith, and R. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," Human Brain Mapping, vol. 2, pp. 189–210, 1995.

[2] R. R. Hocking, Methods and Applications of Linear Models. John Wiley and Sons, 2003.

[3] K. J. Friston, P. Fletcher, O. Josephs, A. Holmes, M. D. Rugg, and R. Turner, "Event-related fMRI: characterizing differential responses." NeuroImage, vol. 7(1), pp. 30–40, 1998.

[4] R. N. Henson, C. J. Price, M. D. Rugg, R. Turner, and K. J. Friston, "Detecting latency differences in event-related BOLD responses: application to words versus nonwords and initial versus repeated face presentations," NeuroImage, vol. 15(1), pp. 83–97, 2002.

[5] V. D. Calhoun, M. C. Stevens, G. D. Pearlson, and K. A. Kiehl, "fMRI analysis with the general linear model: removal of latency-induced amplitude bias by incorporation of hemodynamic derivative terms," NeuroImage, vol. 22(1), pp. 252–257, 2004.

[6] K. J. Worsley and J. E. Taylor, "Detecting fMRI activation allowing for unknown latency of the hemodynamic response," NeuroImage, vol. 29(2), pp. 649–654, 2006.

[7] A. J. Schwarz, B. Whitcher, A. Gozzi, T. Reese, and A. Bifone, "Study-level wavelet cluster analysis and data-driven signal models in pharmacological MRI." J Neurosci Methods, vol. 159(2), pp. 346–360, 2007.

[8] M. A. Lindquist and T. D. Wager, "Validity and Power in Hemodynamic Response Modeling: A Comparison Study and a New Approach," Human Brain Mapping, vol. 28, pp. 764–784, 2007.

[9] M. Woolrich, T. Behrens, and S. Smith, "Constrained linear basis sets

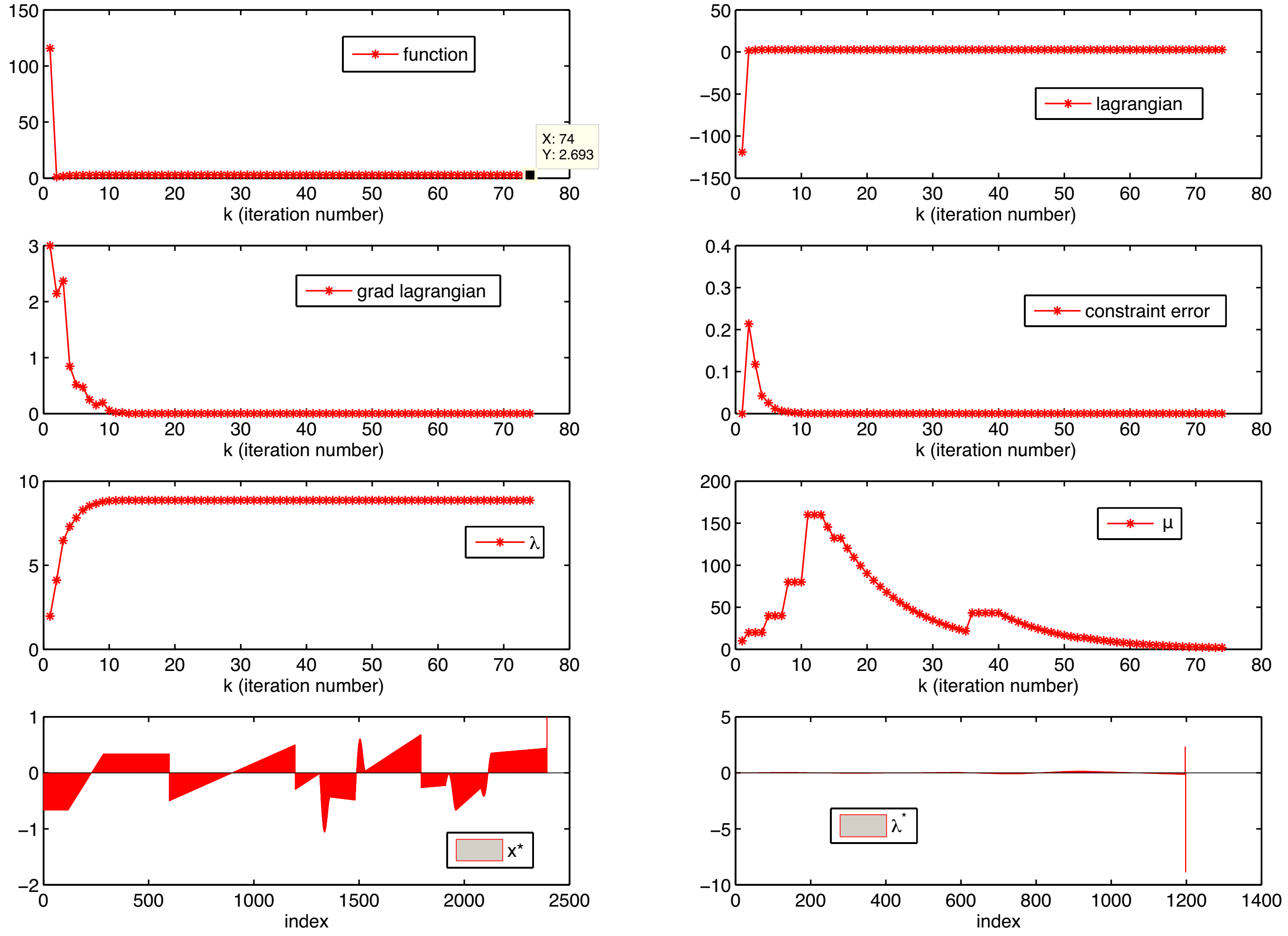
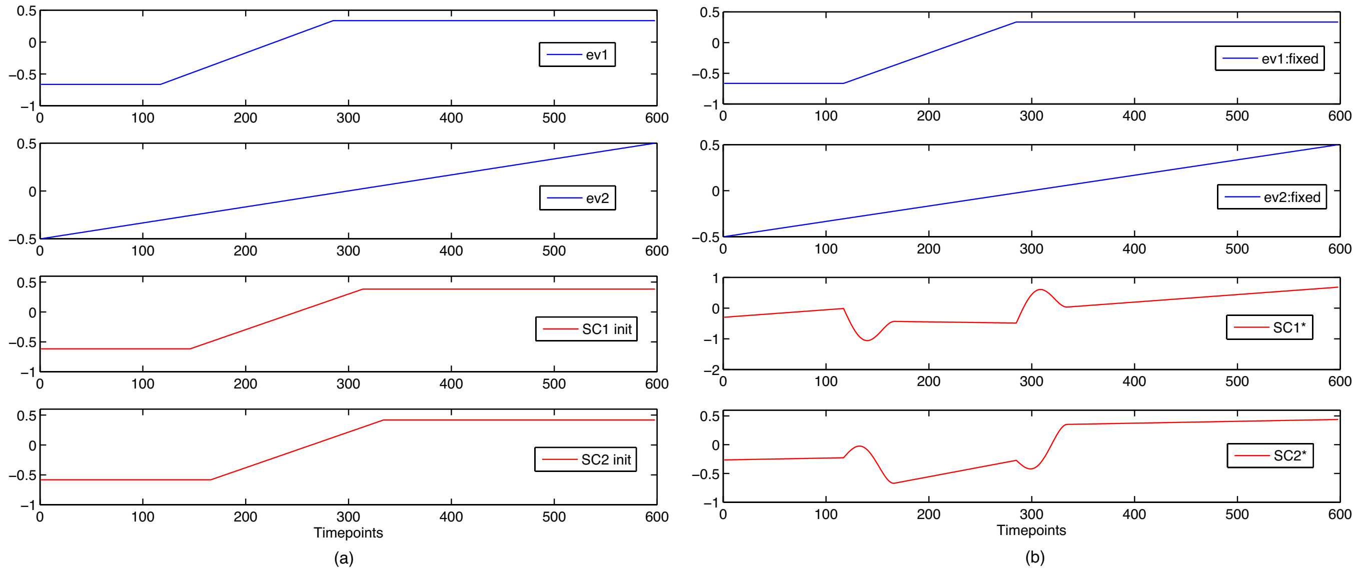


Fig. 31. Validation Test A: Convergence diagnostics for the advanced solver. Figure shows the evolution of objective function, the Lagrangian, norm of the gradient of the Lagrangian, norm of the constraint satisfaction error, norm of the Lagrange multipliers and progress monitoring parameter over algorithm iterations. The last row shows the optimal solution (i.e., the DM and contrast) displayed as a vector and the optimal Lagrange multipliers for the chosen constraints on the columns of Z and the contrast c_Z. The optimal objective of $G(\hat{Z}, \hat{c_Z}) = 2.693$ was attained in 74 iterations.



(a)

(b)

Fig. 32. Validation Test A: (a) Initial design matrix Z_0 and (b) Optimal design matrix $\hat{Z}$ ($m = 50$). The first two EVs were constrained to their shift 0 values. The contrast $\hat{c_Z}$ was constrained to be $[1, 0, 0, 0]^T$.

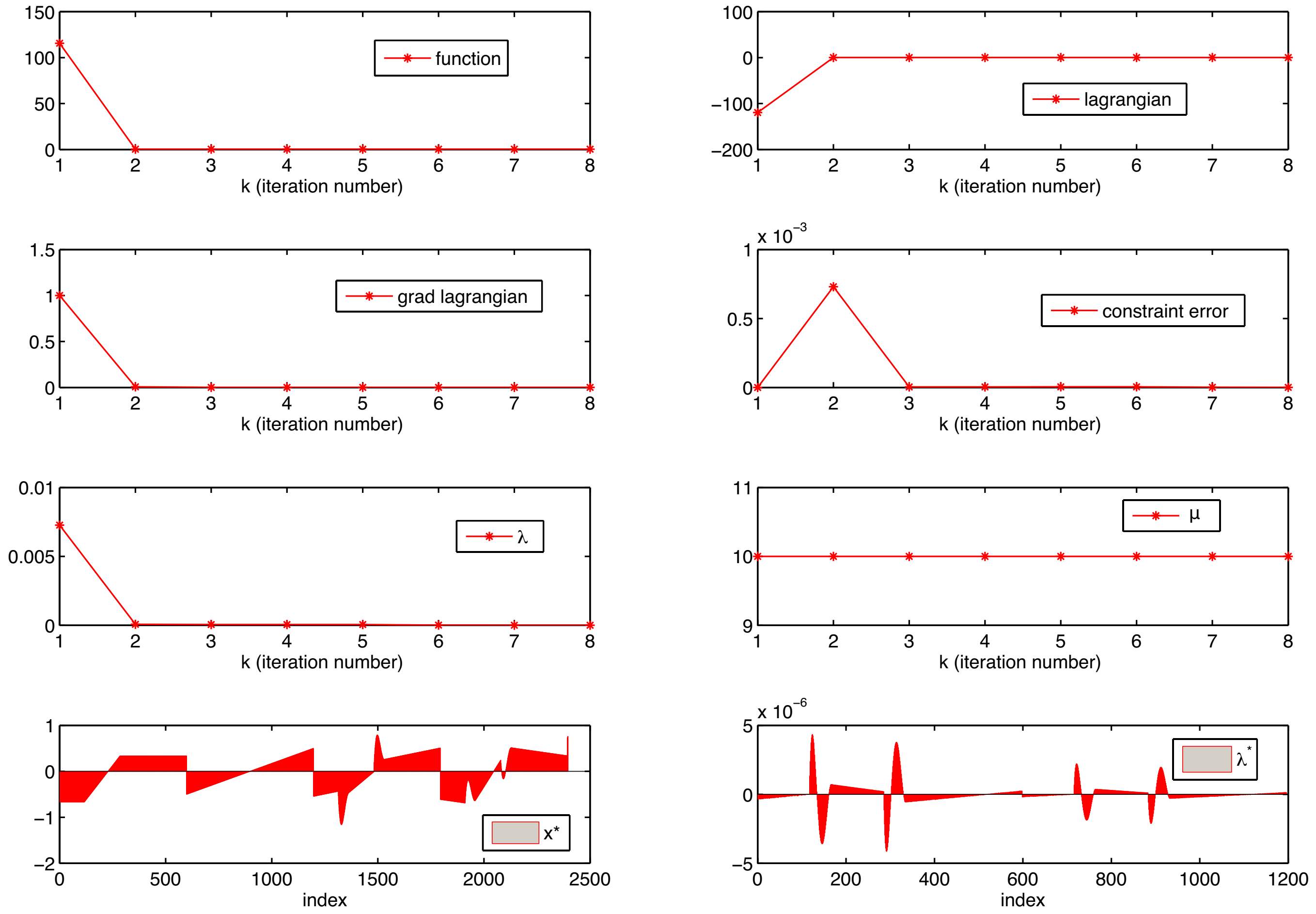


Fig. 33. Validation Test B: Convergence diagnostics for the advanced solver. Figure shows the evolution of objective function, the Lagrangian, norm of the gradient of the Lagrangian, norm of the constraint satisfaction error, norm of the Lagrange multipliers and progress monitoring parameter over algorithm iterations. The last row shows the optimal solution (i.e., the DM and contrast) displayed as a vector and the optimal Lagrange multipliers for the chosen constraints on the columns of Z and the contrast c_Z. The optimal objective of $G(\hat{Z}, \hat{c_Z}) = 0.2655$ was attained in 8 iterations.

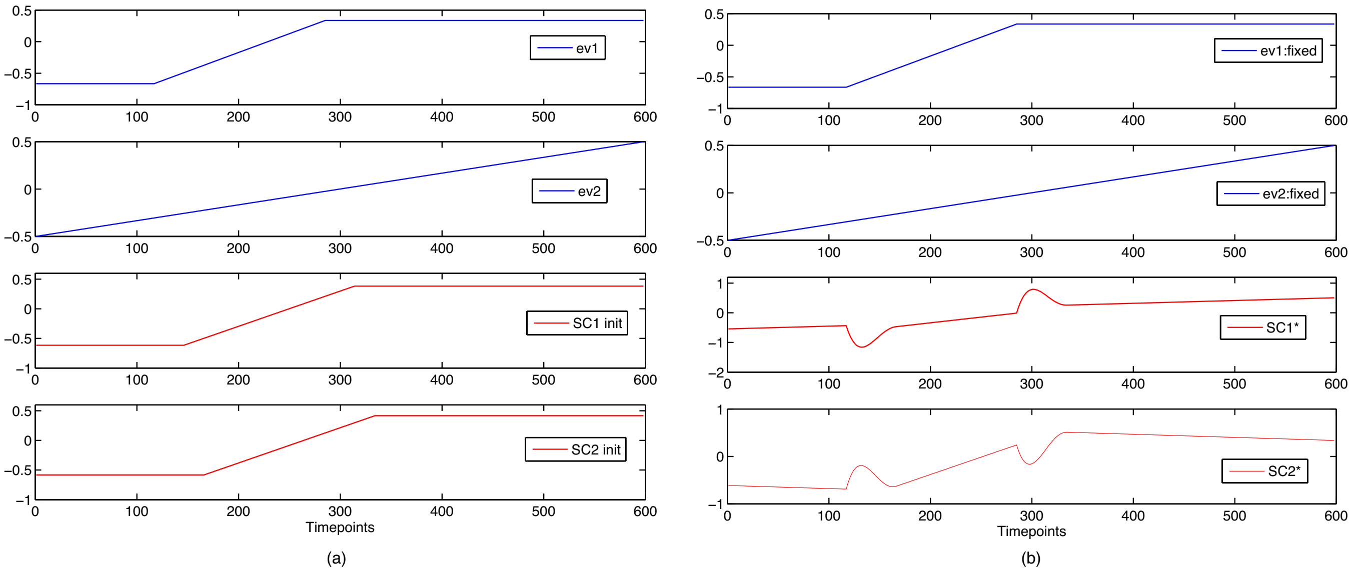
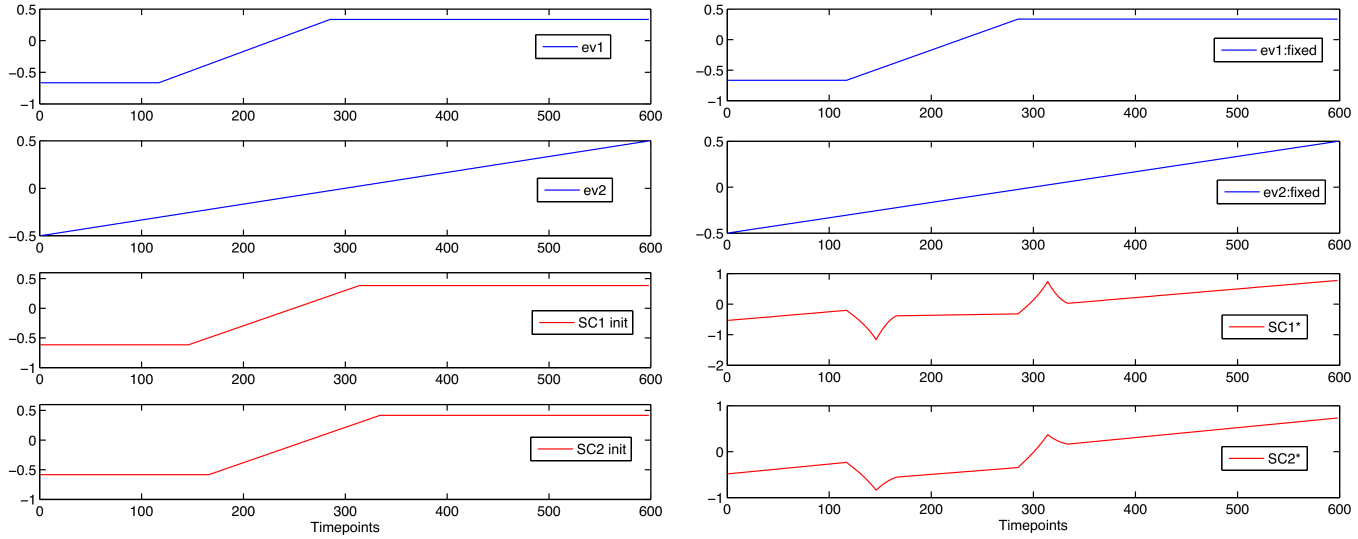
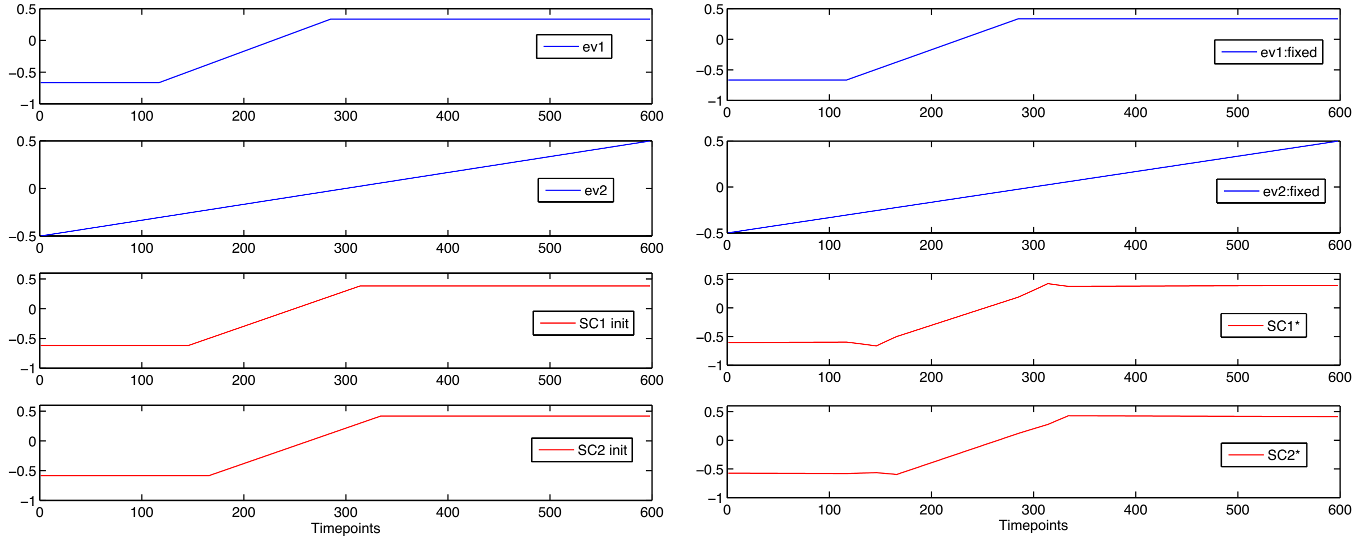


Fig. 34. Validation Test B: (a) Initial design matrix Z_0 and (b) Optimal design matrix $\hat{Z}$ $(m = 50)$. The first two EVs were constrained to their shift 0 values. The contrast $\hat{c_Z}$ was left unconstrained.



(a) (b)

Fig. 35. Validation Test A: (a) Initial design matrix Z_0 and (b) Optimal design matrix $\hat{Z}$ ($m = 50$). The first two EVs were constrained to their shift 0 values. The contrast $\hat{c_Z}$ was constrained to be $[1; 0; 0; 0]$. The problem was solved using **Algorithm 1**. The optimal objective was $G(\hat{Z}, \hat{c_Z}) = 2.693490$.



(a) (b)

Fig. 36. Validation Test B: (a) Initial design matrix Z_0 and (b) Optimal design matrix $\hat{Z}$ ($m = 50$). The first two EVs were constrained to their shift 0 values. The contrast $\hat{c_Z}$ was left unconstrained. The problem was solved using **Algorithm 1**. The optimal objective was $G(\hat{Z}, \hat{c_Z}) = 0.265502$

for HRF modelling using Variational Bayes." NeuroImage, vol. 21(4), pp. 1748–1761, 2004.

[10] J. M. Loh, M. A. Lindquist, and T. D. Wager, "Residual analysis for detecting mis-modeling in fMRI," Statistica Sinica, vol. 18, pp. 1421–1448, 2008.

[11] N. Kriegeskorte, W. K. Simmons, P. S. F. Bellgowan, and C. I. Baker, "Circular analysis in systems neuroscience: the dangers of double dipping," Nature Neuroscience, vol. 12, pp. 535–540, 2009.

[12] T. Nichols and S. Hayasaka, "Controlling the familywise error rate in functional neuroimaging: a comparative review," Statistical Methods in Medical Research, vol. 12, pp. 419–446, 2003.

[13] E. T. Bullmore, J. Suckling, S. Overmeyer, S. Rabe-Hesketh, E. Taylor, and M. J. Brammer, "Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain," IEEE Transactions on Medical Imaging, vol. 18(1), pp. 32–42, 1999.

[14] S. Hayasaka and T. E. Nichols, "Validating cluster size inference: random field and permutation methods," Neuroimage, vol. 20(4), pp. 2343–2356, 2003.

[15] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: a primer with examples." Human Brain Mapping, vol. 15(1), pp. 1–25, 2002.

[16] A. R. Conn, N. I. M. Gould, and P. L. Toint, "A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds," SIAM J. Numerical Analysis, vol. 28, pp. 545–572, 1991.

[17] A. R. Conn, N. I. M. Gould, and P. Toint, "LANCELOT: A Fortan package for large-scale nonlinear optimization (Release A)," Springer Series in Computational Mathematics, vol. 17, 1992.

[18] J. More and D. Sorensen, "Computing a trust region step," SIAM J. Sci. Stat. Comp., vol. 4, pp. 533–572, 1983.

[19] T. Steihaug, "The conjugate gradient method and trust regions in large scale optimization," SIAM J. Numerical Analysis, vol. 20, pp. 626–637, 1983.

[20] R. H. Byrd, P. Lu, J. Nocedal, and C. Y. Zhu, "A limited memory algorithm for bound constrained optimization," SIAM Journal on Scientific Computing, vol. 16, no. 6, pp. 1190–1208, 1995.

[21] A. R. Conn, N. I. M. Gould, and P. L. Toint, "Convergence of quasi-

Algorithm 4: F1: Outer Iteration

```
Require: Initial point $x_{init}$, $\lambda^0$, $\mu_0$, $\theta^h \in (1,\infty)$, $\theta_l \in (0,1)$
Choose tolerances $\eta^*_{con}$ and $\eta^*_{grad}$. The default is $\eta^*_{con} = \eta^*_{grad} = 1e{-}6$. .
$\mu = \mu_0$, $\eta_{con} = 1/\mu_0^{0.1}$, $\eta_{grad} = 1/\mu_0$
for $k = 0, 1, 2, \dots$ do
    $found = 0$
    while $found \neq 1$ do
        Try to find $x_k$ such that $||x_k - P(x_k - \nabla_x \mathcal{L}(x_k, \lambda^k, \mu_k), l, u)||_\infty \leq \eta_{grad}$ via F2 using starting point as $x_{k-1}$.
        if above step is completed successfully then
            Set $found = 1$
        else
            $\lambda^{k+1} = \lambda_k$
            $\mu_{k+1} = \theta_l \mu_k$
            $\eta_{con} = 1/\mu_k^{0.1}$
            $\eta_{grad} = 1/\mu_k$
        end if
    end while
    if $||c(x_k)||_\infty \leq \eta_{con}$ then
        if $||c(x_k)||_\infty \leq \eta^*_{con}$ and $||x_k - P(x_k - \nabla_x \mathcal{L}(x_k, \lambda^k, 0), l, u)||_\infty \leq \eta^*_{grad}$ then
            Stop and return current solution $x_k$.
        end if
        $\lambda^{k+1} = \lambda_k - \mu_k c(x_k)$
        $\mu_{k+1} = \mu_k$
        $\eta_{con} = \eta_{con}/\mu_{k+1}^{0.9}$
        $\eta_{grad} = \eta_{grad}/\mu_{k+1}$
    else
        $\lambda^{k+1} = \lambda_k$
        $\mu_{k+1} = \theta_h \mu_k$
        $\eta_{con} = 1/\mu_k^{0.1}$
        $\eta_{grad} = 1/\mu_k$
    end if
end for
```

Fig. 37. F1: Outer Iteration

Newton matrices generated by the Symmetric Rank One update," Math. Programming, vol. 50, pp. 177–196, 1991.

[22] J. Nocedal and S. Wright, Numerical Optimization, 2nd Edition. New York:Springer, 2006.

[23] C. G. Broyden, "The convergence of a class of double-rank minimization algorithms," J. Ins. Math. Applcs., vol. 6, pp. 76–90, 1970.

[24] E. D. Dolan, J. J. More, and T. S. Munson, "Benchmarking Optimization Software with COPS3," Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, Tech. Rep. ANL/MCS-TM-273, Feb. 2004.

[25] J. Nocedal, "Updating quasi-Newton matrices with limited storage," Math. Computation, vol. 35, pp. 773–782, 1980.

Algorithm 5: F2: Inner Iteration

```
Require: $j_{max}$, $\eta_{grad}$, $\Delta$, $l$, $u$, $\lambda^k$, $\mu_k$, $\eta \in (0,1)$, $flag$
$found = 0$
$x = x_{k-1}$, $j = 1$
Compute, $g = \nabla_x \mathcal{L}(x, \lambda^k, \mu_k)$
Estimate $B = \nabla^2_{xx} \mathcal{L}(x, \lambda^k, \mu_k)$ using BFGS, SR1 or limited memory BFGS, limited memory SR1 quasi Newton Updates.
while $found \neq 1$ and $j \leq j_{max}$ do
    Calculate the Cauchy point $p_c$ for problem:
        $\min_p \frac{1}{2} p^T B p + g^T p$   (143)
        s.t. $l - x \leq p \leq u - x$   (144)
        s.t. $||p||_\infty \leq \Delta$   (145)
    using non-linear gradient projection and calculate the current active set $\mathcal{A}$. Let $e_i$ be the unit vector with $1$ at position $i$ and zeros elsewhere. If $i_1, i_2, \dots i_q \notin \mathcal{A}$ then let $\tilde{Q} = [e_{i_1}, e_{i_2}, \dots, e_{i_q}]$.
    $\tilde{g} = \tilde{Q}^T(g + B\,p_c)$ and $\tilde{B} = \tilde{Q}^T B \tilde{Q}$
    Compute the approximate solution $\hat{v}$ to the problem
        $\min_v \frac{1}{2} v^T \tilde{B} v + \tilde{g}^T v$   (146)
        s.t. $l - x \leq p_c + \tilde{Q} v \leq u - x$   (147)
        s.t. $||p_c + \tilde{Q} v||_\infty \leq \Delta$   (148)
    using truncated conjugate gradient iteration (Newton-CG, Steihaug). If $flag = 1$ use preconditioned Newton-CG using the inexact-modified Cholesky factorization.
    Compute $\hat{p} = p_c + \tilde{Q}\hat{v}$
    Calculate $\delta_{\mathcal{L}} = \mathcal{L}(x) - \mathcal{L}(x + \hat{p})$, $\delta_m = 0.5\hat{p}^T B \hat{p} + g^T \hat{p}$ and $\rho = \delta_{\mathcal{L}}/\delta_m$
    if $\rho > \eta$ then
        $x = x + \hat{p}$
    end if
    Compute new trust region radius $\Delta$ using Framework F3.
    Compute, $g = \nabla_x \mathcal{L}(x, \lambda^k, \mu_k)$ if $\rho > \eta$ holds otherwise use the previous value.
    Estimate $B = \nabla^2_{xx} \mathcal{L}(x, \lambda^k, \mu_k)$ using BFGS, SR1 or limited memory BFGS, limited memory SR1 quasi Newton Updates. Do the update even if $\rho < \eta$.
    if $||x - P(x - \nabla_x \mathcal{L}(x, \lambda^k, \mu_k), l, u)||_\infty \leq \eta_{grad}$ then
        $found = 1$
    end if
    $j = j + 1$
end while
```

Fig. 38. F2: Inner Iteration

Algorithm 6: F3:Trust Region Update

```
Require: $\rho$, $\hat{p}$, $\Delta$
if $\rho > 0.75$ then
    if $||\hat{p}||_\infty \leq 0.8\Delta$ then
        $\Delta = \Delta$
    else
        $\Delta = 2\Delta$
    end if
end if
if $0.1 \leq \rho \leq 0.75$ then
    $\Delta = \Delta$
else
    $\Delta = 0.5\Delta$
end if
return $\Delta$
```

Fig. 39. F3:Trust Region Update

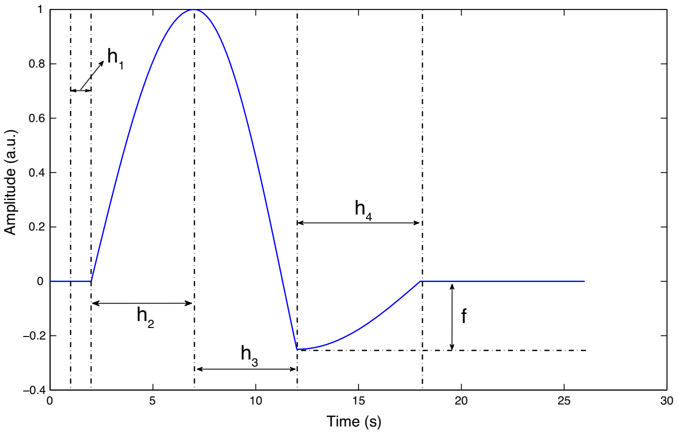


Fig. 40. 5 parameter half-cosine parameterization of Haemodynamic Response Function (HRF). h_1 controls the time to first rise, h_2 controls the time to peak, h_3 controls the time to undershoot maximum and h_4 controls the time to return to baseline. The amplitude of undershoot is controlled by the parameter f while the height of rise from baseline is fixed at 1. See 149 for the exact equation.