University of Nottingham

School of Mathematical Sciences

Sparse regression methods with measurement-error for Magnetoencephalography

Jonathan Samuel Davies

A thesis submitted to the University of Nottingham for the degree of

Doctor of Philosophy

November 2017

Abstract

Magnetoencephalography (MEG) is a neuroimaging method for mapping brain activity based on magnetic field recordings. The inverse problem associated with MEG is severely ill-posed and is complicated by the presence of high collinearity in the forward (leadfield) matrix. This means that accurate source localisation can be challenging. The most commonly used methods for solving the MEG problem do not employ sparsity to help reduce the dimensions of the problem. In this thesis we review a number of the sparse regression methods that are widely used in statistics, as well as some more recent methods, and assess their performance in the context of MEG data. Due to the complexity of the forward model in MEG, the presence of measurement-error in the leadfield matrix can create issues in the spatial resolution of the data. Therefore we investigate the impact of measurement-error on sparse regression methods as well as how we can correct for it. We adapt the conditional score and simulation extrapolation (SIMEX) methods for use with sparse regression methods and build on an existing corrected lasso method to cover the elastic net penalty. These methods are demonstrated using a number of simulations for different types of measurement-error and are also tested with real MEG data. The measurement-error methods perform well in simulations, including high dimensional examples, where they are able to correct for attenuation bias in the true covariates. However the extent of their correction is much more restricted in the more complex MEG data where covariates are highly correlated and there is uncertainty over the distribution of the error.

Acknowledgements

I would like to begin by thanking my supervisors, Dr. Chris Brignell and Prof. Ian Dryden, whose advice and direction has been invaluable. Thank you for your insight, accessibility and for helping to make my PhD studies so enjoyable. I have learned so much under your guidance.

From the Sir Peter Mansfield Magnetic Imaging Centre, I would like to thank Dr. Matt Brookes and Dr. George O'Neill for being available to answer my questions about MEG, providing real data for my analysis and for giving me access to code for leadfield and beamformer calculations. I would also thank the Engineering and Physical Sciences Research Council (EPSRC) for funding my PhD and making my work possible.

Finally, thanks to all my friends and family who have provided a constant source of encouragement and support over the duration of my studies. Your thoughts and prayers have been greatly appreciated. Particular thanks go to my dad for his wisdom, enthusiastic interest in my work and for inspiring me in so many ways.

CONTENTS

1

INTRODUCTION TO MEG

This thesis looks at the statistical methods and models used in various aspects of a neuroimaging technique known as Magnetoencephalography (MEG) (Vrba, 2002; Herdman and Cheyne, 2009; Brookes et al., 2014). Whereas the well known procedure of fMRI measures the blood oxygenation levels of the area being scanned, MEG focuses on the small magnetic fields resulting from the naturally occurring electrical currents during brain activity. By recording the brain's magnetic fields at a number of reference points around the head over a given time interval, the desire is to then reconstruct the activity and more particularly, establish a localisation of active brain areas. The recovered time courses from active areas then allow us to study the temporal behaviour of the activity at various intervals in relation to some stimulus or action. One of the reasons why source localisation is so important is that MEG has a temporal resolution that cannot be matched by other neuroimaging techniques (Papanicolaou, 1998). Currently fMRI (function magnetic resonance imaging) has the advantage in terms of spatial resolution, however it is unable to compete with the sub-millisecond temporal resolution of MEG. Furthermore, MEG directly measures the actual activity itself rather than the secondary effects of brain activity such as blood oxygenation levels. This obviously makes MEG very use-

ful for stimulus-response type experiments as it allows researchers to determine which brain areas are important in certain tasks. However it also allows the possibility of investigating the complex connectivity of neural activity.

The primary challenge with MEG is localising the sources of activity. A common approach is to estimate the strength of activity in each voxel of a 3D grid of voxels of width 8mm using current MEG machines which have nearly 300 sensor locations. This means that the number of potential locations of activity is substantially greater than the number of recording locations. For example, in the grid system described above, the head coordinate model will have many thousands of points. Therefore, it follows that the model represents a severely ill-posed inverse problem (see Hadamard, 1952). Constructing an estimate for the solution to the inverse problem requires a formulation of the forward solution in the form of the leadfield matrix. This requires knowledge of the position at which we are estimating relative to the external sensors and how the magnetic field from a theoretical source with given strength and orientation will behave in the head before it is recorded at each of the sensors. We model the sources as equivalent current dipoles (ECD) which group together the current from many neurons with the same direction as one dipole. A current dipole is a current with location, direction and magnitude which gives off a magnetic field according to the right hand rule. We represent current dipoles as vectors with a strength and orientation.

1.1 WHAT IS MEG?

Magnetoencephalogy (MEG) is a non-invasive neuroimaging method that measures the naturally occurring magnetic fields that result from synaptic activity

in the brain. The brain's magnetic activity is considerably weaker than that of environmental noise, other human biomagnetic rhythms and even the earth's magnetic field, therefore MEG requires both magnetic shielding of the environment and incredibly sensitive detectors. The most commonly used MEG systems deploy an array of SQUIDs (superconducting quantum interference devices) that broadly follow the contours of the scalp in order to measure the disruption in magnetic fields at various positions around the head. This recorded activity can often be attributed to the pyramidal cells in the cortex that are oriented tangentially to the scalp. These sources of activity are represented as current dipoles, with a location, orientation and magnitude. The orientation of these sources with respect to the scalp is important as, following the right hand law, the magnetic field of a dipole points around its direction (see Fig. 2). In other words, a sensor that is radially oriented to a dipole will be unable to recover the magnetic field.

Ultimately, the goal in MEG is to use the sensor recordings of induced magnetic fields from the SQUID array in order to reconstruct the activity within the brain itself. This requires us to be able to solve the inverse problem (see section 1.3).

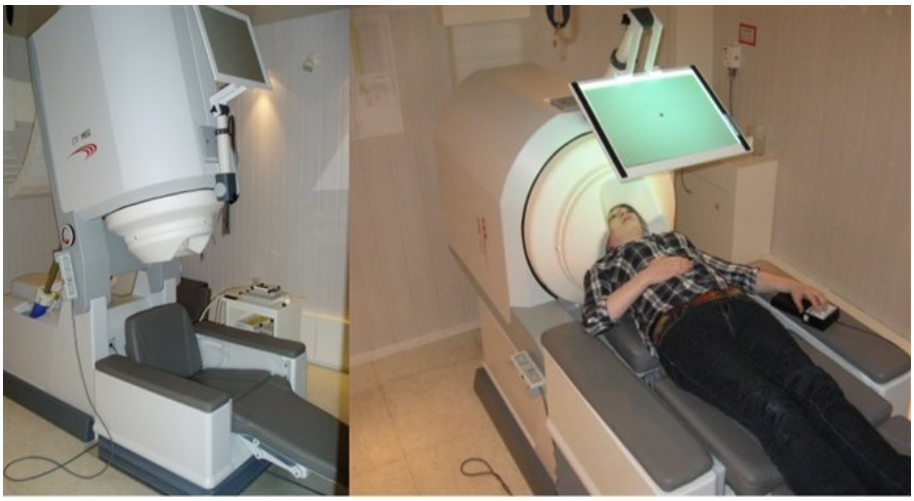

Figure 1: MEG scanner at Sir Peter Mansfield Magnetic Imaging Centre (SPMMIC), University of Nottingham. Patient undergoing an experiment, scanner can be used in horizontal or vertical position. Source: `https://www.nottingham.ac.uk/magres/facilities/meg.aspx`

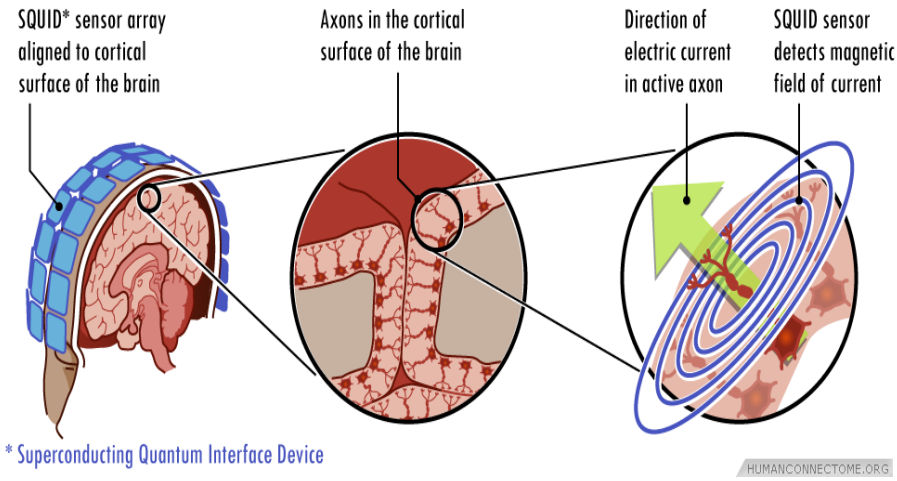

Figure 2: Representation of brain activity as current dipole. Magnetic field given by source in relation to scanner location. Source: `https://www.humanconnectome.org/about/project/MEG-and-EEG.html`

1.2 MEG EXPERIMENTS

MEG is primarily used to research brain functionality and connectivity. Consequently, MEG experimental paradigms typically present the subject with some stimulus from which we look to investigate the change in the recorded activity over its duration. This activity generally falls into one of two categories; evoked or induced responses. Broadly, evoked responses are phase locked and so we can expect some 'average' activity to exist over multiple repetitions of the same experiment. Conversely, induced responses have no temporal phase locking and so any averaged response will likely flatten out the activity. The type of activity will dictate the approach that we take to processing the data (Salmelin and Parkkonen, 2010).

Images of the recorded activity can be produced by integrating information from magnetic resonance imaging (MRI). During the recording session of the experiment, the subject will have a set of fiducial points (typically the nasion and the left/right pre-auricular) marked with electrified coils. These points can then be co-registered with similar markers in the MRI image. Note, the MRI image is recorded in a separate session and with a different machine, so it is important that the markers are set up correctly. From this we are able to represent the head as a 3-dimensional object made up of appropriately sized voxels. Each of these voxels then represents a location at which we want to estimate the neuronal activity. From the markers we are also able to set up a coordinate system so that the voxels in the head model correspond with particular areas in the MRI images, and hence we can interpret the results in terms of the head anatomy.

If we recall from the representation of sources as current dipoles, the orientation of the source is also important for determining the measured magnetic field. Therefore MEG modelling also incorporates the source's orientation into the estimation. This is represented through the coordinate system of the head model. For each voxel we can construct a plane tangent to the location vector, over which we can optimise to find the orientation (see Fig. 3). Note that the orientation is optimised on a two dimensional plane due to the insensitivity to radial contributions of the source. In practice, the orientation is usually determined in one of two ways; either we estimate the source for each of 180 angles and then choose the direction that maximises the source power, or we estimate in two orthogonal directions from which we are able to extract the magnitude and orientation information.

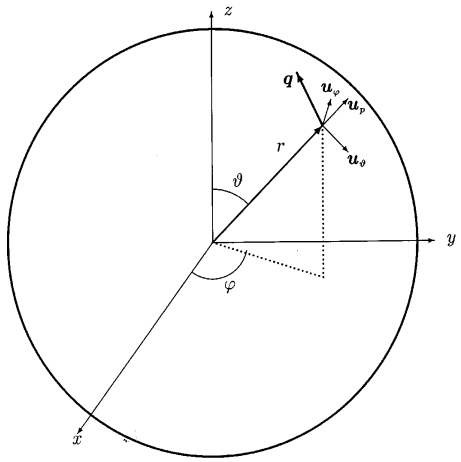

Figure 3: Representation of coordinate system in the head. Vector r is the position of the source in the system. The tangential plane over which we optimise the orientation of $\boldsymbol{q}$ is defined by the unit vectors $\boldsymbol{u}_\theta, \boldsymbol{u}_\psi$. Note in MEG we are unable to obtain the radial contribution of the source, so we ignore the radial unit vector $\boldsymbol{u}_p$. (Adapted from image in Dogandzic and Nehorai, 2000).

This approach, where we reconstruct the activity at each location in our coordinate grid is known as the distributed source approach. The alternative

current dipole framework models the current as a small set of sources. These sources are represented as dipoles with location, orientation and strength parameters. Under this approach the model is now over-determined, however in practice, the estimation of the source parameters is often more complex than estimating the entire vector field since the model is non-linearly related to the dipole locations (Mosher et al., 1992). In addition, the number of sources must be estimated from the data. Methods under this approach make common use of Bayesian schemes for likelihood maximisation (e.g. Kiebel et al., 2008). An important consideration in dipole modelling is how to allow the activity to evolve over time. For example, it is preferable to keep the number of dipoles small, but allow new dipoles to appear and existing dipoles to disappear. Rather than allow movement of dipoles, the neurophysiology dictates that dipole locations should remain fixed. The static dipole method (Sorrentino et al., 2013) models the evolution of the dipoles over time through a transition density, which includes the possibility of a new dipole appearing, an existing dipole disappearing and remaining dipoles moving in moment (strength and orientation).

There are advantages for both distributed source and current dipole models. The former can be modelled linearly and as such are usually easier to solve, however dipole models are often able to explain the data through a small number of parameters and consequently, can be easier to interpret. In this thesis though, we will primarily focus on distributed source approaches for MEG source localisation.

1.3 MEG MODEL

The MEG model for n sensors, T time points and p locations/orientations of interest, is as follows:

$$\mathbf{d} = \mathbf{Ls}$$

where $\mathbf{d}$ is an $n \times T$ matrix of the measured potentials outside the head, $\mathbf{L}$ is the $n \times p$ matrix of the forward solutions (leadfield matrix) where the i, j^{th} element represents the recorded activity at the i^{th} sensor resulting from a unit source at the j^{th} location/orientation, and the $p \times T$ source matrix $\mathbf{s}$ is the current amplitude from the source in p orientations and locations at each time point (Hauk, 2004). Given our recorded $\mathbf{d}$ and the leadfield determined by the location of the source and the spherical head model, we want to estimate $\mathbf{s}$, the source amplitude. In reality, the recordings will contain some error $\boldsymbol{\epsilon}$ which represents noise and outside interference. So the model is given by:

$$\mathbf{d} = \mathbf{Ls} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$. Since the magnetic fields in the brain are very small, this error becomes more dominant the further we go into the brain itself. This is represented by the fact that the variance of the leadfield columns corresponding to deeper, more central, locations is smaller than those related to shallower regions. This means that, assuming the sensor noise is not related to the source location, the signal-to-noise ratio reduces as we go further into the brain.

1.4 LEADFIELDS

The estimation of source activity in MEG studies is based on the forward solution contained in the matrix $\mathbf{L}$. This matrix is known as the "leadfield" matrix and each column $\mathbf{L}_\theta$ represents the magnetic field that is measured at each sensor as a response to a unit dipole at location/orientation θ. In practice, for a given dipole with known location and orientation, the calculation of the leadfield matrix involves solving Maxwell's equations to give the resulting magnetic fields (Mosher et al., 1999). We can then determine the magnetic potentials that would be measured at a number of sensor locations. Each column of the leadfield matrix then represents the values of the magnetic potential at a number of sensors resulting from a single dipole. An example of a leadfield matrix for a single location with 180 orientations and 270 sensors is found in Fig. 4. As we can see, at any particular sensor, the forward solutions across the orientations form a smooth curve.

Obviously, the accuracy of the estimation is hugely dependent on how well we are able to model the theoretical recordings of a source. In reality, the forward solution is quite complex and requires knowledge of biomagnetism, the behaviour of magnetic fields and their permeability through different tissues in the human head. We also need some idea of the geometry of the head. There are a number of different ways that we can bring the geometry of the head into the calculation of the leadfields, but a common approach is to approximate the shape of the head using spheres. Obviously a single sphere is a far too simplistic approximation for a shape as complex (and varied) as the human head, however an alternative is to use multiple overlapping local spheres (Lalancette et al., 2011). This approach fits an optimal sphere for each sensor in the array such

that the sensor is oriented tangentially to the chosen sphere. Although the multiple spheres approach is a vast improvement on using a single spherical head model, there is still some element of abstraction. This opens up the possibility of some error being introduced into the model. Additionally, since the leadfield model assumes a fixed position of the head, any head movement by the subject in the scanner will be problematic. Despite the best efforts of experimenters to reduce head movement (Meyer et al., 2017), movement of the head will result in spatial blurring and localisation errors (Stolk et al., 2013). Therefore, there is some motivation for investigating the effects of measurement-error in the leadfield matrix on MEG data.

We will now introduce the two methods that are most commonly used to solve the MEG inverse problem, namely the beamformer and minimum norm.

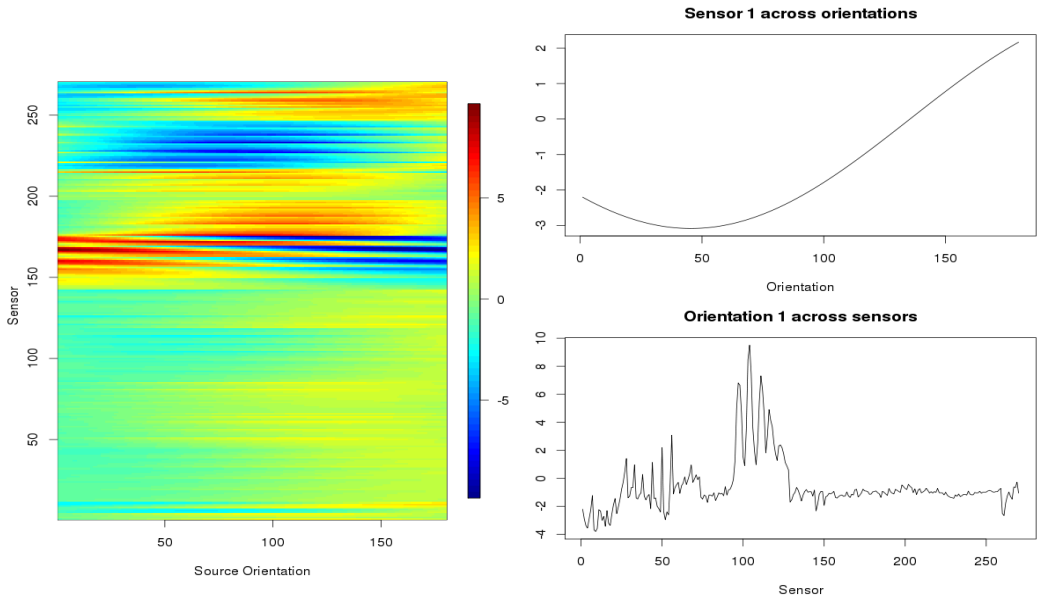

Figure 4: Typical leadfield for a location; Image of leadfield for 270 sensors and 180 orientations, plot for first sensor, plot for first orientation (left, top right, bottom right).

1.5 MINIMUM NORM

The minimum norm (Hamalainen and Ilmoniemi, 1984; Wang et al., 1993; Dale and Sereno, 1993) looks to produce a linear estimator $\mathbf{G}$ such that we can produce an estimate of the source amplitude,

$$\hat{\mathbf{s}} = \mathbf{Gd} = \mathbf{GLs} = \mathbf{Rs}$$

where the matrix $\mathbf{R}$ can be seen as the distortion of the true measured source amplitude. Ideally, we would be able to find the estimate of source amplitude $\hat{\mathbf{s}}$ such that $\mathbf{L}\hat{\mathbf{s}} = \mathbf{d}$. However, in reality this will not often be the case and instead we will look to minimise the squared difference between the LHS and RHS. i.e at each time point $i = 1, 2, ..., T.$,

$$\min_{\hat{\mathbf{s}}_i}(\mathbf{d}_i - \mathbf{L}\hat{\mathbf{s}}_i)^T(\mathbf{d}_i - \mathbf{L}\hat{\mathbf{s}}_i) = \min_{\hat{\mathbf{s}}_i}||\mathbf{d}_i - \mathbf{L}\hat{\mathbf{s}}_i||_2^2$$

or equivalently find $\hat{\mathbf{s}}_i$ such that,

$$\min_{\hat{\mathbf{s}}_i}\|\mathbf{d}_i - \mathbf{L}\hat{\mathbf{s}}_i\|_2.$$

Hence, this is known as the “minimum norm" approach. The solution to this is then found using standard techniques and is given by,

$$\hat{\mathbf{s}} = (\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T\mathbf{d}. \tag{1.2}$$

Note this is equivalent to the least squares solution. This solution corresponds to the case when we have no prior knowledge about the distribution. In reality, the matrix $\mathbf{L}^T\mathbf{L}$ is singular and regularisation will be required to invert it, i.e

$$\hat{\mathbf{s}} = (\mathbf{L}^T\mathbf{L} + \lambda\mathbf{I})^{-1}\mathbf{L}^T\mathbf{d}$$

where λ is the regularisation parameter and $\mathbf{I}$ is the identity matrix. Hence $\hat{\mathbf{s}}$ is the same as the ridge regression estimator (Hoerl and Kennard, 1970a,b), also

known as Tikhonov regularisation (Tikhonov, 1943). This unconstrained form of the minimum norm will only be of real use if our p locations/orientations of estimation includes the true location of the source. In other locations, it will attempt to fit the data to the location of interest and will give a solution that does not really fit the problem. A more general form of the minimum norm can be gained by incorporating weighting matrices into the solution. First a matrix inversion lemma is applied to equation (1.2) to give,

$$\hat{\mathbf{s}} = \mathbf{L}^T(\mathbf{L}\mathbf{L}^T)^{-1}\mathbf{d}$$

(Sekihara and Nagarajan, 2015). Then taking $\mathbf{C}_s$ to be a weight matrix representing our prior knowledge of the source locations and covariances, and $\mathbf{C}_r$ to be a sensor noise covariance matrix, then we have a general form of the weighted minimum norm,

$$\hat{\mathbf{s}} = \mathbf{C}_s^{-1}\mathbf{L}^T\left(\mathbf{L}{\mathbf{C}_s}^{-1}\mathbf{L}^T + \lambda\mathbf{C}_r\right)^{-1}\mathbf{d} \tag{1.3}$$

(Hauk, 2004). Setting the weighting matrices both to the identity matrix and thereby applying equal regard to all sensors and source locations, we arrive at the previous form of the minimum norm.

1.6 BEAMFORMER

One of the popular methods of solving the MEG model is beamforming. Beamforming is a technique that was originally applied to radar data and was adapted for brain data (Van Veen et al., 1997; Robinson and Vrba, 1998). The formulation that follows is based on the outline of beamforming in Brookes et al. (2008). Rather than solve equation (1.1) by a standard regression approach, beamforming produces an estimate of the source amplitude at some location and orienta-

tion. We take $\theta = (r, \delta)$ to represent a particular location/orientation where r is the location of interest and δ is an orientation on the plane given by $\boldsymbol{u}_\theta, \boldsymbol{u}_\psi$ as in Fig. 3. The beamformer estimate for a given θ, denoted $\hat{\mathbf{Q}}_\theta$, therefore represents an estimate for the corresponding row in $\mathbf{s}$ from equation (1.1), and is based on a weighted sum of the measured sensor outputs. i.e.

$$\hat{\mathbf{Q}}_\theta = {\mathbf{w}_\theta}^T \mathbf{d}, \tag{1.4}$$

where the recorded sensor data $\mathbf{d}$ is mean centred over time. Proceeding, a subscript θ denotes the row or column of a matrix that corresponds to a given location/orientation θ.

The beamforming approach of a weighted sum can also be seen to be a spatial filter of the recorded data with the filter specified by $\mathbf{w}_\theta$. In order to obtain the beamformer weight for a particular location $\mathbf{w}_\theta$, we minimise the expected source power with the constraint that the power from location/orientation θ remains in the estimate i.e.

$$\mathbb{E}\left[\hat{\mathbf{Q}}_\theta^2\right] = {\mathbf{w}_\theta}^T \mathbb{E}\left[\mathbf{d}\mathbf{d}^T\right] \mathbf{w}_\theta$$

subject to ${\mathbf{w}_\theta}^T \mathbf{L}_\theta = 1$. Since the $\mathbf{L}_\theta$'s represent the forward solution given a unit dipole for the particular location/orientation θ the weighted sum for θ must be equal to one. However, where possible, we want to minimise the weighted sums for different locations/orientations i.e block the contributions of other sources. Mathematically, we would ideally want the matrix $\mathbf{W}^T\mathbf{L}$ that is constructed from the individual weighted sums to be equal to the order n identity matrix, however in reality we have a matrix that is only restricted to have 1's on the lead diagonal. We minimise the above using the method of Lagrange multipliers as follows: If we write $\mathbf{C} = \mathbb{E}\left[\mathbf{d}\mathbf{d}^T\right]$ and with our linear constraint set,

$$f\left(\mathbf{w}_\theta, \lambda\right) = {\mathbf{w}_\theta}^T \mathbf{C} \mathbf{w}_\theta + \lambda\left(1 - {\mathbf{w}_\theta}^T \mathbf{L}_\theta\right).$$

Then differentiating with respect to $\mathbf{w}_\theta$ and setting equal to zero,

$$\frac{\partial f}{\partial \mathbf{w}_\theta} = 2\mathbf{w}_\theta{}^T\mathbf{C} - \lambda\mathbf{L}_\theta{}^T = 0,$$

$$\Rightarrow \hat{\mathbf{w}}_\theta^T = \frac{\lambda}{2}\mathbf{L}_\theta{}^T\mathbf{C}^{-1}. \tag{1.5}$$

Now differentiating with respect to λ and setting equal to zero,

$$\frac{\partial f}{\partial \lambda} = 1 - \mathbf{w}_\theta{}^T\mathbf{L}_\theta = 0,$$

$$\Rightarrow \mathbf{w}_\theta{}^T\mathbf{L}_\theta = 1.$$

Then substituting (1.5) in for $\mathbf{w}_\theta^T$,

$$\frac{\lambda}{2}\mathbf{L}_\theta{}^T\mathbf{C}^{-1}\mathbf{L}_\theta = 1$$

$$\Rightarrow \hat{\lambda} = 2\left(\mathbf{L}_\theta{}^T\mathbf{C}^{-1}\mathbf{L}_\theta\right)^{-1}.$$

Finally, substituting $\hat{\lambda}$ back into (1.5) we obtain,

$$\hat{\mathbf{w}}_\theta^T = \left(\mathbf{L}_\theta{}^T\mathbf{C}^{-1}\mathbf{L}_\theta\right)^{-1}\mathbf{L}_\theta{}^T\mathbf{C}^{-1}. \tag{1.6}$$

Hence $\hat{\mathbf{Q}}_\theta$ is a weighted least squares estimate with weight matrix $\mathbf{C}$ and overall we have,

$$\hat{\mathbf{Q}}_\theta = \left(\mathbf{L}_\theta{}^T\mathbf{C}^{-1}\mathbf{L}_\theta\right)^{-1}\mathbf{L}_\theta{}^T\mathbf{C}^{-1}\mathbf{d} = \hat{\mathbf{w}}_\theta^T\mathbf{d}.$$

Note that if $\mathbb{E}(\mathbf{d}) = \mathbf{0}$ then the beamformer is the same as a generalised least squares where $\mathrm{Var}(\mathbf{d}) = \mathbf{C}$. Hence, it is important that the data $\mathbf{d}$ is mean centred over time.

1.6.1 *Weights normalisation*

One issue with MEG is that the signal-to-noise ratio reduces as we go deeper into the brain. Consequently, at the centre of the brain, the noise power begins

to dominate that of the genuine source activity. In order to account for the depth of the location of interest and the effect it has on the beamformer estimate it is common for the estimate to be divided by some normalisation factor. One such method, known as weights normalisation is to divide the beamformer estimate by $D = \sqrt{\mathbf{w}_\theta^T \mathbf{w}_\theta}$, which is chosen to give unbiased noise variance estimates across all the voxels in the model (Luckhoo et al., 2014). Using the weights derived above,

$$\begin{aligned} D &= \left[\left(\mathbf{L}_\theta{}^T \mathbf{C}^{-1} \mathbf{L}_\theta \right)^{-1} \mathbf{L}_\theta{}^T \mathbf{C}^{-1} \mathbf{C}^{-1} \mathbf{L}_\theta \left(\mathbf{L}_\theta{}^T \mathbf{C}^{-1} \mathbf{L}_\theta \right)^{-1} \right]^{\frac{1}{2}} \\ &= \left(\mathbf{L}_\theta{}^T \mathbf{C}^{-1} \mathbf{L}_\theta \right)^{-1} \sqrt{\mathbf{L}_\theta{}^T \mathbf{C}^{-2} \mathbf{L}_\theta} \end{aligned}$$

since $\left(\mathbf{L}_\theta{}^T \mathbf{C}^{-1} \mathbf{L}_\theta \right)^{-1}$ is just a scalar. Then dividing our beamformer estimate by D we get,

$$\hat{\mathbf{Q}}_\theta^{wn} = \frac{\mathbf{L}_\theta{}^T \mathbf{C}^{-1} \mathbf{d}}{\sqrt{\mathbf{L}_\theta{}^T \mathbf{C}^{-2} \mathbf{L}_\theta}}. \tag{1.7}$$

Similarly, we can produce a noise normalised form of the minimum norm. Recalling the weighted minimum norm estimate from equation 1.3, let $\hat{\mathbf{s}}_{mn} = \hat{\mathbf{w}}_{mn}\mathbf{d}$, where $\hat{\mathbf{w}}_{mn} = \mathbf{C}_s^{-1}\mathbf{L}^T \left(\mathbf{L}\mathbf{C}_s{}^{-1}\mathbf{L}^T + \lambda \mathbf{C}_r \right)^{-1}$. Then for the k^{th} location/orientation, the noise-normalised minimum norm estimate is,

$$\hat{\mathbf{s}}_{nn-mn,k} = \frac{\hat{\mathbf{s}}_{mn,k}}{w_k}, \tag{1.8}$$

where w_k is a weighting constant. A common choice of w_k is the dynamic statistical parametric mapping (dSPM), where $w_k^2 = (\hat{\mathbf{w}}_{mn}\mathbf{C}_r\hat{\mathbf{w}}_{mn}^T)_{kk}$ (Dale et al., 2000). Hence, when $\mathbf{C}_r = \mathbf{I}$ the weighting is similar to the beamformer weight D. An alternative noise-normalised minimum norm estimate is obtained using the sLORETA weight, $w_k^2 = (\hat{\mathbf{w}}_{mn}(\mathbf{C}_r + \mathbf{L}\mathbf{C}_s{}^{-1}\mathbf{L}^T)\hat{\mathbf{w}}_{mn}^T)_{kk}$ (Pascual-Marqui, 2002). Statistical maps such as the dSPM and sLORETA introduced above represent a common framework for brain imaging. There are a number of different approaches to statistical mapping of brain activity. While many

of these methods will regard each time point as independent, there are also approaches that look to include temporal information. For example Kilner and Friston (2010) used random fields theory to produce a mapping that is continuous in time.

Additionally, for the standard beamformer, we can produce a whole head image of the electrical activity using the estimated source power. Let P_θ be the source power for location/orientation θ, then,

$$P_\theta = \mathbb{E}\left[\hat{\mathbf{Q}}_\theta^2\right] = \mathbf{w}_\theta^T \mathbf{C} \mathbf{w}_\theta = \left(\mathbf{L}_\theta{}^T \mathbf{C}^{-1} \mathbf{L}_\theta\right)^{-1}$$

but again, we require some normalisation in order to account for the depth. In this case we use an estimate of the noise power at the given location, represented by $\rho_\theta = \mathbf{w}_\theta^T \mathbf{\Sigma}_\epsilon \mathbf{w}_\theta$, where $\mathbf{\Sigma}_\epsilon$ is the channel noise variance. Therefore ρ_θ has the same form as the source power estimate. Assuming the noise is uncorrelated and approximately equal across channels, we can use $\mathbf{\Sigma}_\epsilon = \sigma^2 \mathbf{I}$ where σ^2 is the noise variance and $\mathbf{I}$ is the n dimensional identity matrix. Normalising the source power we have what is known as the "pseudo Z statistic",

$$Z_\theta = \frac{P_\theta}{\rho_\theta} = \frac{\mathbf{w}_\theta^T \mathbf{C} \mathbf{w}_\theta}{\mathbf{w}_\theta^T \mathbf{\Sigma}_\epsilon \mathbf{w}_\theta}.$$

The pseudo Z statistic is also known as the neural activity index and can be calculated at each location in the brain in order to give an image of brain activity. At each location r, we can use the neural activity index to determine the orientation of the source activity. Since MEG is insensitive to radial sources, the optimisation of the source orientation reduces to determining the angle on a plane tangential to the radial direction. To do this we calculate,

$$Z_{opt} = \max_\delta Z_{r,\delta}, \quad 0^\circ \leq \delta \leq 180^\circ.$$

In practice, we calculate Z_θ for each of our orientations at location r and then choose the δ that gives the maximum value. For some datasets, we may want

to measure the change in power during some task. In particular, if we have an active and control period we might want to determine the significance of any power change between the periods. In a similar way to the pseudo Z-statistic, we can construct a "pseudo T-statistic",

$$T_\theta = \frac{P_\theta^{(act)} - P_\theta^{(con)}}{\rho_\theta^{(act)} + \rho_\theta^{(con)}}$$

where the superscripts $(act), (con)$ refer to the active and control periods respectively. Naturally, this requires that a separate covariance matrix be calculated for each window.

1.6.2 *Covariance matrix*

If we assume that the forward solution is accurate, then the accuracy of the beamformer depends entirely on the estimate of the covariance matrix $\mathbf{C}$. For a single source $\mathbf{Q}_1$ with forward solution $\mathbf{L}_1$, analytically the covariance matrix is

$$\begin{aligned}\mathbf{C} =& \mathbb{E}\left(\mathbf{d}\mathbf{d}^T\right) = \mathbb{E}\left((\mathbf{L}_1\mathbf{Q}_1 + \sigma\boldsymbol{\epsilon})(\mathbf{L}_1\mathbf{Q}_1 + \sigma\boldsymbol{\epsilon})^T\right) \\ =& \mathbb{E}\left(\mathbf{L}_1\mathbf{Q}_1\mathbf{Q}_1^T\mathbf{L}_1^T + \sigma^2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T + 2\sigma\mathbf{L}_1\mathbf{Q}_1\boldsymbol{\epsilon}^T\right) \\ =& \mathbf{Q}_1^2\mathbf{L}_1\mathbf{L}_1^T + \sigma^2\mathbf{I}.\end{aligned}$$

In reality, the matrix $\mathbf{C}$ is not known and it therefore must be estimated from the data $\mathbf{d}_1, ..., \mathbf{d}_T$ over time. We estimate $\mathbf{C}$ using $\mathbf{C}_d$ as follows for the time centred $\mathbf{d}_i$'s. Each entry $\mathrm{C}_{d,(i,j)}$ is the following sum,

$$\mathbf{C}_{d,(i,j)} = \frac{1}{T}\sum_{t=1}^{T} d_i(t)d_j(t)$$

where $t = 1, 2, ..., T$ is the time step. The covariance matrix is an important part of the beamformer as the beamformer estimate and the source power estimate

are actually based on the variance of the source rather than its strength. The accuracy of the covariance matrix $\mathbf{C}_d$, can be studied as follows. Firstly, we can write the data at time t as,

$$\mathbf{d}(t) = \mathbf{d}_0(t) + \boldsymbol{\epsilon}(t)$$

where $\mathbf{d}_0(t)$ is the true measurement of the magnetic field at time t, and $\boldsymbol{\epsilon}(t)$ is the sensor noise. Substituting into the previous expression for the covariance we have

$$\mathbf{C}_{d,(i,j)} = \frac{1}{T}\sum_{t=1}^{T}(d_{0,i}(t) + \epsilon_i(t))(d_{0,j}(t) + \epsilon_j(t)).$$

Expanding out we have,

$$\begin{aligned}
\mathbf{C}_{d,(i,j)} &= \frac{1}{T}\sum_{t=1}^{T} d_{0,i}(t)d_{0,j}(t) + \frac{1}{T}\sum_{t=1}^{T} d_{0,i}(t)\epsilon_j(t) + \frac{1}{T}\sum_{t=1}^{T} \epsilon_i(t)d_{0,j}(t) \\
&\quad + \frac{1}{T}\sum_{t=1}^{T} \epsilon_i(t)\epsilon_j(t) \\
&= \frac{1}{T}\sum_{t=1}^{T} d_{0,i}(t)d_{0,j}(t) + c1 + c2 + c3 \\
&= \mathbf{C}_{0,(i,j)} + \Delta\mathbf{C}_{i,j}
\end{aligned}$$

where $\Delta\mathbf{C} = c1 + c2 + c3$ is the error in the covariance matrix and $\mathbf{C}_0$ is the noiseless data covariance matrix (Brookes et al., 2008). Over an infinite limit of integration, the error $\Delta\mathbf{C}_{i,j}$ will tend towards zero. In finite integration limits, the error will be finite and $c1, c2$ are uncorrelated, however the pairs $(c1, c3)$ and $(c2, c3)$ will be weakly correlated as long as $\boldsymbol{\epsilon}$ does not dominate $\mathbf{d}_0$. If this is the case, then an upper bound on the variance of $\Delta\mathbf{C}_{i,j}$ can be expressed as, $\mathrm{Var}(\Delta\mathbf{C}_{i,j}) = \mathrm{Var}(c1) + \mathrm{Var}(c2) + \mathrm{Var}(c3)$ where,

$$\mathrm{sd}(c1) = \frac{\mathrm{sd}(d_{0,i})\sigma}{\sqrt{T}}$$

$$\mathrm{sd}(c2) = \frac{\mathrm{sd}(d_{0,j})\sigma}{\sqrt{T}}$$

$$\mathrm{sd}(c3) = \begin{cases} \frac{\sigma^2}{\sqrt{T}} & \text{if } i \neq j \\ \sqrt{\frac{2}{T}}\sigma^2 & \text{if } i = j \end{cases}$$

Hence the variance of the random process can be expressed,

$$\mathrm{Var}(\Delta \mathbf{C}_{i,j}) = \frac{\sigma^2}{T}\{\mathrm{sd}(d_{0,i})^2 + \mathrm{sd}(d_{0,j})^2 + [1 + \mathbb{1}(i = j)]\sigma^2\}.$$

1.7 MEG EXAMPLE

To demonstrate the performance of the minimum norm and beamformer for some typical MEG data we present a simple MEG simulation. For 100 trials, each of 500 time points, a single source was simulated with each trial consisting of a variation of a damped cosine waveform (with random shifting and scaling). Normally distributed noise was added to the data with mean 0 and variance 100. The minimum norm was performed on the data averaged over trials and the beamformer was applied to the full data (some discussion on trial averaging and the differing approaches for the minimum norm and beamformer can be found in section 2.2).

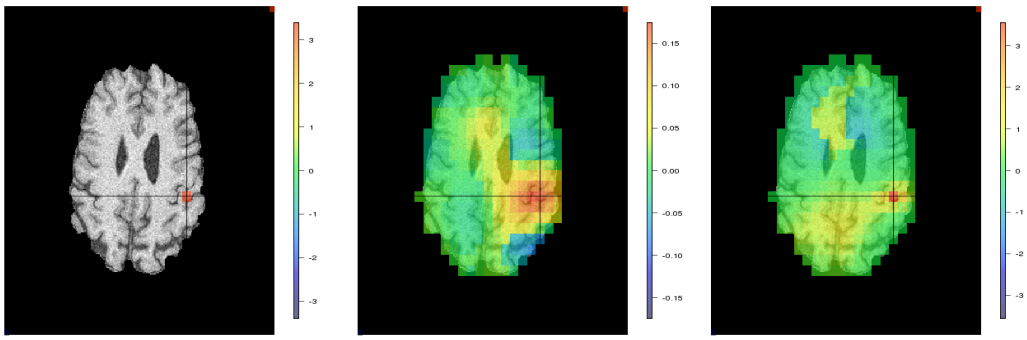

Figure 5: (l-r) True source, minimum norm, beamformer.

Fig.5 shows the truth, along with the minimum norm and beamformer estimates when the source is at its maximum strength. The minimum norm esti-

mate is shown to spread out the activity of the source to nearby locations, an issue known as signal leakage (Wens et al., 2015). The peak in the beamformer estimate is much sharper and there is good recovery of the source strength and location. However, despite the good performance of the beamformer at the true source location, weak activity is also placed throughout the rest of the brain. Furthermore, when correlated sources are present, the beamformer often has trouble reconstructing the activity correctly (Brookes et al., 2007). A commonality in both the minimum norm and the beamformer is that neither method is able to set covariates corresponding to inactive locations to be zero. A common approach to high dimensional problems in statistics is to reduce the number of predictors involved in the model through the use of sparse models. Sparse models are able to employ selection techniques to identify unimportant covariates that can be dropped from the model. This is usually done through an ℓ_1 penalty, however use of the traditional ℓ_1 penalised lasso in MEG has been shown to produce estimates that are spatially unstable and temporally spiky (Uutela et al., 1999). Due to these issues there have been a number of approaches that have looked to apply different penalisations to the spatial and temporal aspects of MEG data. For example, Ou et al. (2009) approached the problem using a group norm and Tian et al. (2012) proposed a two way regularisation method that iteratively optimises an ℓ_1 penalty on the spatial aspect of the problem and a smoothing penalty on the temporal dimension. Similarly, Solin et al. (2016) used a Gaussian process model to regularise the solution in space and time. For the beamformer, Zhang et al. (2014) proposed a covariance thresholding approach that also introduces sparsity into the estimates.

Many of these methods require us to simultaneously (or recursively) optimise over the spatial and temporal dimensions, which can be quite demanding. How-

ever there are a number of extensions to the standard lasso that have not been applied to MEG. These methods usually combine the ℓ_1 and ℓ_2 penalties in order to improve the stability of the estimates in the presence of highly correlated predictors.

1.8 THESIS OUTLINE

The outline of this thesis is as follows, in Chapter 2 we review a number of sparse regression methods and discuss their application in the context of MEG data, before assessing their performance in some real data in comparison to the minimum norm and beamformer methods. Since MEG estimation is based primarily on the model leadfield matrix, it is of interest to investigate measurement-error in the MEG model context. The measurement-error approach will be used in an attempt to allow for an incorrect leadfield matrix. This error in the leadfield could arise from modelling errors or through experimental factors such as head movement and incorrect co-registration (Lopez et al., 2012; Akalin Acar and Makeig, 2013). Therefore, in Chapter 3 the concept of measurement-error is introduced along with the conditional score correction method and the effect of measurement-error on MEG source estimation is investigated. The conditional score is then adapted for use with penalised regression methods in Chapter 4. Continuing with measurement-error, Chapter 5 focuses on the simulation extrapolation (SIMEX) method applied to sparse methods for additive and multiplicative error, and in Chapter 6 we extend a corrected lasso estimate to cover the elastic net method along with results for the theoretical performance.

The measurement-error corrections for sparse methods are compared in the context of Monte Carlo simulations in Chapter 7, and finally they are applied to real data and final conclusions are made in Chapter 8.

2

SPARSE MODELS

INTRODUCTION

Magnetoencephalography (MEG) data presents a number of difficulties in terms of source localisation. The primary complication is the dimensionality of the data. In MEG problems the number of potential locations of interest is far greater than the number of observations at each time point, which is typically around 275 (Vrba, 2002). Furthermore there is frequently a high presence of collinearity amongst predictors due to the spatial proximity of the locations of interest represented in the leadfield matrix. These characteristics make traditional regression techniques either computationally infeasible or at least fairly unsuitable (Hoerl and Kennard, 1970b). A fairly modern statistical approach to such ill-posed problems is to use sparse regression techniques so that a number of the coefficients are set to be zero. The advantage of the sparse regression approach is that we get a form of dimension reduction via variable selection (Tibshirani, 1996). These methods typically require the optimisation of a penalisation term controlling the sparsity, however a suitable choice of penalty parameter will significantly reduce the model complexity and from a theoretical point of view they seem to be appropriate for MEG data.

Despite this, sparse regression methods have seen limited implementation to the MEG problem due to the non-convex nature of the ℓ_1 norm and the comparative slowness in optimisation compared to the ℓ_2 norm (Uutela et al., 1999, has some implementation of ℓ_1 norm regression, Ou et al., 2009). Some work has been done on producing sparse estimates for MEG data, but this has mostly been approached from a Bayesian perspective with sparsity introduced through modelling of the source distribution via components of the covariance matrix. For example the multiple sparse priors method (MSP) (Friston et al., 2008) aims to model the covariance matrix of the true sources by a linear combination of different proposed component covariance matrices that are controlled by hyperparameters. Each of these covariance components represents the covariance resulting from a particular patch in the cortex. The hyperparameters have priors applied to them and they are then updated to convergence using Restricted Maximum Likelihood (REML). Sparsity is therefore introduced by the removal of hyperparameters near zero through the REML process.

Due to the severely ill-posed nature of the MEG inverse problem, the empirical Bayesian approach can be computationally intensive. Specifying a large number of hyperparameters adds extra complexity to the model and consequently the search for an optimal solution can be slow to converge using realistic leadfields (Ramirez et al., 2007). On the other hand, there has been little use of the linear regression based sparse methods described here for source localisation. These methods do not require any Bayesian prior modelling and some of the methods have particularly good qualities with regards to noise invariance (Belloni et al., 2011). In this thesis we will focus on these non-Bayesian methods.

In this chapter, we introduce a number of sparse regression methods that have been applied to similar high dimensional ill-posed problems, along with a more recently developed method, before applying them to basic MEG simulations and then moving on to some typical real data.

2.1 SPARSITY IN THE REGRESSION MODEL

We begin with the standard linear regression model, for n observations and p predictors,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon} \tag{2.1}$$

where the data vector $\mathbf{y}$ has length n, the unknown parameter vector $\boldsymbol{\beta}$ is length p, the model matrix $\mathbf{X}$ has dimensions $n \times p$, σ^2 is some unknown noise variance and $\boldsymbol{\epsilon}$ follows a multivariate standard normal distribution. We want to estimate the parameter vector $\boldsymbol{\beta}$. Assuming the errors $\boldsymbol{\epsilon}$ have conditional mean zero, are spherical (with no homoscedasticity or autocorrelation) and the matrix $\mathbf{X}$ is of full column rank, under the ordinary least squares (OLS) solution, we minimise the sum of squares of the residuals for this model which leads to the well known least squares estimate,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}. \tag{2.2}$$

For low dimensional problems this estimator can be implemented without major issues and will return unbiased, efficient and consistent estimates (Rao, 1973). However high dimensional problems, especially those where $n < p$, prove somewhat problematic. The main issue with the OLS in these situations is that the matrix $\mathbf{X^T X}$ is no longer invertible. This means that the OLS estimator is computationally infeasible without some modification. One way to solve this issue is by employing methods of selection, whereby the number of parame-

ters is reduced to only those that are important. Another approach is to use shrinkage, where parameters are shrunk towards zero by penalising the sum of squares of the residuals. Ridge regression is one example of shrinkage.

2.1.1 *Ridge regression*

We have seen that when the matrix $\mathbf{X^T X}$ is singular (or near singular) the OLS solution does not exist (or is unstable). These problems can be addressed by the introduction of the ridge penalty. Ridge regression is a shrinkage method that penalises the sum of the squared beta estimates (Hoerl and Kennard, 1970a,b). The level of penalisation determines the amount of shrinkage displayed in the model fitting. So, for example a small ridge penalty will allow larger coefficients as it is less restrictive on their size. Mathematically, ridge regression can be summarised as follows,

$$\hat{\boldsymbol{\beta}}_{RR} = \arg\min_{\beta \in \mathbb{R}^p} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2^2. \tag{2.3}$$

where λ is a regularisation constant. The estimator can then be shown to be

$$\hat{\boldsymbol{\beta}}_{RR} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.4}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix. With the addition of the penalty, the matrix $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ is now invertible and the regression coefficients shrink towards zero. Of course, the shrinkage employed in ridge regression is never able to set parameter estimates exactly to zero, so for models with large p the dimension problems are not really addressed. The lasso (Least Absolute Shrinkage and Selection Operator, Tibshirani, 1996) and similar *sparse* regression methods however, are able to ensure that some parameters have the value of exactly zero.

Sparse regression methods work by introducing a penalisation into the model as in ridge regression, but rather than constrain the squared coefficients (via the ℓ_2 norm) the *absolute size* of the chosen beta values is usually constrained to be below some value λ instead. The various sparse methods therefore employ some combination of the ℓ_1 norm and the ℓ_2 norm, where the ℓ_1 norm of the beta values is $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|$ and the ℓ_2 norm is $||\boldsymbol{\beta}||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$.

2.1.2 *Considerations for sparse methods*

When dealing with ill-posed problems there are a number of things to consider in order to help inform our choice of sparse models. Below, we will present the estimators for some of the sparse regression methods that we have used, along with some of the theoretical properties of the given methods. Beforehand, it is worth briefly covering some of the more attractive features that we look for in sparse regression methods.

One of the most important things to consider in large dimensional problems is the selection of variables. All the sparse methods essentially look to reduce the number of 'important' variables that are included in a model; it is this property that gives them their name. Of course, the choice of which parameters to set to zero, or more importantly which parameters contribute to the model, is a crucial factor in the performance of a given model. For a sparse model with true parameter vector $\boldsymbol{\beta}_0$ let us define the "active set" S_0 as,

$$S_0 = \{j : \beta_{0,j} \neq 0\},$$

i.e. the set of the non-zero parameters. Now let $s_0 = |S_0|$, where $|.|$ here denotes the cardinality of a set, be the sparsity index and essentially represents the level of sparsity when compared to the dimensions of the full model. Naturally,

the true S_0 is unknown and therefore, the selection of non-zero coefficients in penalised regression looks to estimate this set. A useful property for a sparse method to have is therefore the oracle property (Fan and Li, 2001). The oracle property states that the estimator is asymptotically equivalent to the ideal estimator (under no regularisation) with only the true parameters. In other words, there exists a particular value of the regularisation parameter λ where,

$$\mathbb{P}(\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) \to 1.$$

An alternative way of looking at oracle performance in an estimator is, under suitable choice of regularisation (i.e. $\lambda \to 0$ as $n \to \infty$), the difference between the estimator and the true parameters is below some value that tends to zero as n increases. i.e.

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 = O(n^{-1/2} + a_n),$$

where a_n depends on λ and $a_n \to 0$ with increasing n (Fan and Li, 2001). This is the way that oracle (or near-oracle) performance of a method is often expressed. This property obviously depends on the choice of λ and whilst we can sometimes derive an ideal value of the regularisation from the theory, in practice it is not always trivial to determine the most appropriate theoretical value.

2.1.3 *The lasso*

The lasso (Tibshirani, 1996) differs from ridge regression in the use of the ℓ_1 penalty in place of the ridge's ℓ_2 norm,

$$\hat{\boldsymbol{\beta}}_{Las} = \arg\min_{\beta \in \mathbb{R}^p} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1. \tag{2.5}$$

Therefore, both the lasso and ridge regression seek to find the $\boldsymbol{\beta}$ that minimizes the residual sum of squares (RSS) subject to some penalty. The two norms can be visualised graphically in the case of two dimensions. In ridge regression, the plot for the curve $||\boldsymbol{\beta}||_2 = 1$ results in a circle centred on zero, whereas the plot of $||\boldsymbol{\beta}||_1 = 1$ from the lasso gives a diamond. It is possible therefore to see the sparsity properties of the lasso. The minimising beta solution of the lasso is therefore very likely to lie on one of the corners of the diamond. The corners represent sparse solutions where one (or more in larger dimensions) of the parameters is equal to zero. Some discussion of penalty geometry is given in section 2.1.7. As with all sparse methods, the choice of λ value is very important as each value of λ will give a different $\boldsymbol{\beta}$ solution.

The choice of λ is informed by the theory, more particularly a value that gives a consistent estimator. Under fairly mild assumptions on the errors, for suitable λ it can be shown that,

$$||\mathbf{X}(\hat{\boldsymbol{\beta}}_{Las} - \boldsymbol{\beta}_0)||_2^2/n \leq \frac{3}{2}\lambda||\boldsymbol{\beta}_0||_1$$

and furthermore that λ is of order $\sqrt{\frac{\log(p)}{n}}$ (Bühlmann and van de Geer, 2011). Then the estimator is consistent if $||\boldsymbol{\beta}_0||_1 = o\{\sqrt{n/log(p)}\}$ (Bühlmann and van de Geer, 2011). In order to achieve optimality in the convergence of the parameter vector, we require some design conditions to hold. Examples of these are the restricted eigenvalue assumption (Bickel et al., 2009) and the compatibility condition (van de Geer, 2007; van de Geer and Bühlmann, 2009). When the restricted eigenvalue assumption holds and under normal noise, if we take the regularisation parameter to be

$$\lambda = \sigma c 2n^{1/2}\Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) \tag{2.6}$$

for constant $c > 1$, we have the near-oracle performance of the estimator. That is with probability greater than $1 - \alpha$,

$$||\hat{\boldsymbol{\beta}}_{Las} - \boldsymbol{\beta}_0||_2 \lesssim \sigma\sqrt{s_0 \frac{\log(2p/\alpha)}{n}},$$

(Belloni et al., 2011). The lasso has the benefit of being a convex optimisation problem that can be easily solved computationally through a variety of approaches, however it also has some drawbacks. When the coefficients are large, the lasso has been shown to show some bias. It is also worth noting that the λ used in order to gain near oracle performance requires knowledge of the noise standard deviation σ. In problems when $p >> n$, estimating the noise becomes rather challenging and the performance of the lasso may be affected.

2.1.4 *Elastic net*

The strength of the lasso is its ability to perform both shrinkage and variable selection simultaneously. However some features of the lasso make it unsuitable for certain situations. In situations where $p > n$ the lasso will select at most n variables. Also, the lasso is not well defined (i.e. not unique) unless the ℓ_1 norm bound is below some value. Furthermore, in situations where variables are pairwise highly correlated the lasso will choose one of the correlated variables without distinction (Zou and Hastie, 2005). These properties mean that in problems where the number of parameters is much larger than the number of observations and when many of these parameters are highly correlated with each other, the lasso is no longer equipped to provide the best solutions. The elastic net was devised to overcome these limitations, by taking advantage of the strengths of both the lasso and ridge regression. The elastic net provides a

versatile method that combines the ℓ_1 norm penalty and the ℓ_2 norm penalty from the lasso and ridge regression. The elastic net solution is

$$\hat{\boldsymbol{\beta}}_{EN} = \arg\min_{\beta \in \mathbb{R}^p} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 ||\boldsymbol{\beta}||_2^2. \tag{2.7}$$

Therefore, the elastic net can be seen as a generalisation of both methods. i.e. $\lambda_2 = 0$ gives the special case of the lasso and $\lambda_1 = 0$ gives ridge regression. The challenge of choosing the level of the constraints (i.e. λ_1 and λ_2) is now further complicated by the fact that there are now two lambda values to be chosen. This can be done by pre-specifying some choices of values for the ridge penalty and then solving the lasso path for each of them via a method such as the LARS algorithm (Efron et al., 2004). i.e. Set $\lambda_2 \in \{\lambda_{2,1}, \lambda_{2,2}...., \lambda_{2,K}\}$ and then for each $\lambda_{2,k}$, where $k = 1, 2, ..., K$, solve over the lasso solution path.

An alternative way of formulating the elastic net is to introduce a parameter that defines the ratio between the two penalties. If we take $a = \frac{\lambda_2}{\lambda_1 + \lambda_2}$, then an equivalent optimisation is

$$\hat{\boldsymbol{\beta}}_{EN} = \arg\min_{\beta \in \mathbb{R}^p} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \left\{ (1-a)||\boldsymbol{\beta}||_1 + a||\boldsymbol{\beta}||_2^2 \right\}.$$

The elastic net can be shown to be an augmented version of the lasso optimisation using the following formulation (Zou and Hastie, 2005). Let $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$, $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$, and then construct the data $(\mathbf{y}^*, \mathbf{X}^*)$, where

$$\mathbf{X}^* \frac{1}{\sqrt{1+\lambda_2}} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I} \end{pmatrix}, \quad \mathbf{y}^* = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}.$$

The elastic net problem can then be written in the form of a lasso objective function,

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}^* \in \mathbb{R}^p} ||\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*||_2^2 + \gamma ||\boldsymbol{\beta}^*||_1.$$

Then $\hat{\boldsymbol{\beta}}_{EN} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^*$. It is important to note that the dimensions of the design $\mathbf{X}^*$ are $(n+p) \times p$, and so the matrix has rank p. This means that

the elastic net estimate is able to select more than n important predictors and is no longer constrained by the sample size as in the lasso. Furthermore, due to the strict convexity of the elastic net penalty when $\lambda_2 > 0$, the elastic net possesses a grouping property, namely that the difference between highly correlated predictors will be small. These two points contribute to the good performance of the elastic net in the area of variable selection. In practice, the elastic net often works better if it is close to the lasso or ridge penalties (i.e. a is close to 0 or 1 respectively) and as such the formulation above is sometime known as the "naive elastic net" (Zou and Hastie, 2005). It also seems to suffer from double shrinkage due to the use of two penalties. We can correct for the extra shrinkage in the naive elastic net. Using the augmented data as above, the rescaled elastic net estimate is defined

$$\hat{\boldsymbol{\beta}}_{EN-res} = \sqrt{1+\lambda_2}\hat{\boldsymbol{\beta}}^*,$$

however recall that $\hat{\boldsymbol{\beta}}_{EN} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^*$ and so

$$\hat{\boldsymbol{\beta}}_{EN-res} = (1+\lambda_2)\hat{\boldsymbol{\beta}}_{EN}.$$

Scaling the estimate reduces the effect of shrinkage on the elastic net, whilst preserving the important properties.

2.1.5 *Square root lasso*

The square root lasso (SRL) is a modification of the lasso that looks to minimise the square root of the residual sum of squares subject to the ℓ_1 norm constraint (Belloni et al., 2011). i.e.

$$\hat{\boldsymbol{\beta}}_{SRL} = \arg\min_{\beta\in\mathbb{R}^p} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2 + \frac{\lambda}{n}||\boldsymbol{\beta}||_1. \tag{2.8}$$

The method has the advantage of being *pivotal*, that is it requires no knowledge of the value of the noise variance σ^2. Furthermore, the method remains valid even in situations with non-Gaussian noise as long as $\mathbb{E}(\epsilon) = 0$ and $\mathbb{E}(\epsilon^2) = 1$. The estimation of the noise standard deviation σ which is required in the lasso is non-trivial when the number of parameters is especially large, so by dispensing of the need to estimate σ, the square root lasso has particularly good theoretical properties with respect to the solution of the optimisation problem. The dependence of the lasso on the noise variance estimate can be seen in the presence of σ in the optimal choice of the penalty (see equation 2.6). In the square root lasso, it can be shown that by using the penalty

$$\lambda = cn^{1/2}\Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) \tag{2.9}$$

where the constant $c > 1$ (the choice of $c = 1.1$ is recommended by Belloni et al., 2011), the near oracle performance is returned i.e.

$$||\hat{\boldsymbol{\beta}}_{SRL} - \boldsymbol{\beta}_0||_2 \lesssim \sigma\sqrt{s_0\frac{\log(2p/\alpha)}{n}}.$$

This means that under certain design conditions the square root lasso will attain similar performance to the lasso even without any knowledge of the noise variance. The absence of σ in 2.9 means that the square root lasso is a pivotal method with respect to λ. Additionally, this remains the case under non-Gaussian noise as mentioned earlier. These properties are a great advantage over the standard lasso.

2.1.6 *Penalised Euclidean distance*

Penalised Euclidean distance regression (PED) resembles the elastic net in that it combines the use of both the lasso and ridge penalties (Vasiliu et al., 2014).

However, in contrast to the elastic net, there is only one λ value used. The PED constraint employs a geometric mean of the ℓ_1 and ℓ_2 norms,

$$\hat{\boldsymbol{\beta}}_{PED} = \arg\min_{\beta \in \mathbb{R}^p} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2 + \lambda\sqrt{||\boldsymbol{\beta}||_2.||\boldsymbol{\beta}||_1}. \tag{2.10}$$

The ℓ_2 norm is introduced in order to add some consideration of correlated variables to the useful properties seen in the square root lasso. Therefore, the PED estimate looks to combine the attractive aspects of both the square root lasso and the elastic net. The grouping property of the geometric mean penalisation $\sqrt{||\boldsymbol{\beta}||_2.||\boldsymbol{\beta}||_1}$ comes from the fact that when f, g are norms, $\sqrt{f.g}$ is itself a norm (Vasiliu et al., 2014). Then the grouping properties can be shown as follows. Firstly, identical columns of the design will give estimates that are also equivalent, or mathematically,

$$col_i(\mathbf{X}) = col_j(\mathbf{X}) \Rightarrow \hat{\beta}_{PED,i} = \hat{\beta}_{PED,j}.$$

Secondly, if we define the relative difference between the estimates of two parameters as $D_{(i,j)} = \frac{1}{||\hat{\beta}_{PED}||_2}|\hat{\beta}_{PED,i} - \hat{\beta}_{PED,j}|$ then it can be show that

$$D_{(i,j)} \leq \frac{2\sqrt{1-\rho_{ij}}}{\lambda} \leq \frac{2\theta_{i,j}}{\lambda},$$

where ρ_{ij} is the sample correlation and $0 \leq \theta_{i,j} \leq \frac{\pi}{2}$ is the angle between the i'th and j'th columns of $\mathbf{X}$ (Vasiliu et al., 2014). Hence for highly correlated variables, the difference between the PED estimates will be small. Also, as we introduce more penalisation into the PED model by increasing λ the estimates become more sensitive to any correlation between variables. We can therefore see that the PED method displays a grouping property.

In addition to the grouping property we also have sparsity in the model thanks to the use of the ℓ_1 norm. In situations where the true model is very sparse, the PED method will display similar performance to the square root lasso. This

can be seen by considering the geometric interpretation of the penalties (see next section). An important measure of the sparsity of the PED estimate is,

$$\hat{k} = \sqrt{\frac{||\hat{\boldsymbol{\beta}}_{PED}||_2}{||\hat{\boldsymbol{\beta}}_{PED}||_1}}$$

where $\frac{1}{\sqrt[4]{p}} \leq \hat{k} \leq 1$ as long as $\hat{\boldsymbol{\beta}}_{PED} \neq \mathbf{0}$. The performance of the PED estimate in general can be assessed in terms of its oracle inequality as with the other methods. Under restricted eigenvalue conditions, where κ is one such restricted eigenvalue defined in Vasiliu et al. (2014) and $0 < \rho < 1$, for a suitable choice of λ, the recovered parameter estimates can be shown to satisfy the following oracle inequality,

$$\kappa||\hat{\boldsymbol{\beta}}_{PED} - \boldsymbol{\beta}_0||_2 \leq \frac{const.\sqrt{p^* \log(2p/\alpha)}L(\boldsymbol{\beta}_0)}{n(1-\rho^2)}$$

where $L(\boldsymbol{\beta}_0)$ is the Euclidean loss $||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0||_2$ and p^* is the cardinality of $\boldsymbol{\beta}_0$. So as n grows, the estimate approaches the true parameter values. As a corollary to this main result, let $0 < \xi < 1$, and suppose the restricted eigenvalues under ξ are bounded away from zero, then taking $\lambda = \frac{c\sqrt[4]{p}}{n}\Phi^{-1}\left(1 - \frac{\alpha}{2p}\right)$ with $c > 1$ we can check that,

$$\sqrt{\frac{||\hat{\boldsymbol{\beta}}_{PED}||_2}{||\hat{\boldsymbol{\beta}}_{PED}||_1}} - \frac{\sqrt{n}}{c\sqrt[4]{p}} = \hat{k} - \frac{\sqrt{n}}{c\sqrt[4]{p}} \geq \xi > 0. \tag{2.11}$$

Then with probability $1 - \alpha$ we have the oracle property

$$||\hat{\boldsymbol{\beta}}_{PED} - \boldsymbol{\beta}_0||_2 \leq \frac{const.\sqrt{p^* \log(2p/\alpha)}}{\sqrt{n}}.$$

The theory above can be used to inform our choice of which parameter estimates are useful. The theoretical basis tells us that at parameters that have no signal, $\beta_j = 0$, $|\hat{\beta}_{PED,j}| < ||\hat{\boldsymbol{\beta}}_{PED} - \boldsymbol{\beta}_0||_2 \leq \frac{const.\sqrt{p^* \log(2p/\alpha)}}{\sqrt{n}}$. Then as long as $||\hat{\boldsymbol{\beta}}_{PED}||_2 \neq 0$, we can divide by $||\hat{\boldsymbol{\beta}}_{PED}||_2$ to give,

$$\frac{|\hat{\boldsymbol{\beta}}_{PED,j}|}{||\hat{\boldsymbol{\beta}}_{PED}||_2} < \frac{const.\sqrt{p^* \log(2p/\alpha)}}{\sqrt{n}||\hat{\boldsymbol{\beta}}_{PED}||_2} < \frac{C}{\sqrt{n}},$$

where $C \propto \sqrt{p^* \log(2p/\alpha)} > 0$ is a constant that can be determined by AIC/BIC or chosen to provide a certain sparsity level. This inequality is used to produce a threshold by which we infer which β_j's are irrelevant. Parameters with estimates below this threshold value will simply be dropped from the model and the optimisation problem is considered again with the remaining parameters, hence sparsity is introduced into the model. Evidently, the choice of the thresholding constant C becomes very important as increasing its value will reduce the number of parameters chosen. It is often a case of finding a value that works for the problem of interest through some information criterion or similar method. A sequence of values is specified for both C and λ. Then a grid search is employed to find the combination of C, λ that gives the smallest value of BIC.

Computationally, the PED algorithm involves performing convex optimisation on the objective function with some initial value of λ before coefficients that fall below the threshold are removed. This is then repeated for a sequence of λ values and the best fitting model is chosen using BIC or another criteria. Finally, using the λ from the best model, the optimisation is repeated using only the parameters that were chosen beforehand. An alternative to using BIC to choose the best value of λ is to employ the theoretical oracle properties from the corollary. Equation 2.11 motivates the maximisation of the sparsity measure $\hat{k}$ as a method for choosing the parameters. We will refer to this maximisation of $\hat{k}$ as the ξ method.

2.1.7 *Geometry of penalties*

We now briefly discuss the geometric interpretation of the various penalties. The geometry can be useful for understanding why a particular penalty promotes certain behaviour in the estimation. The plots below give an indication of the behaviour of the different penalties by looking at the 2 dimensional case. Fig. 6 shows the contour plot for the various penalties for the unit penalty level.

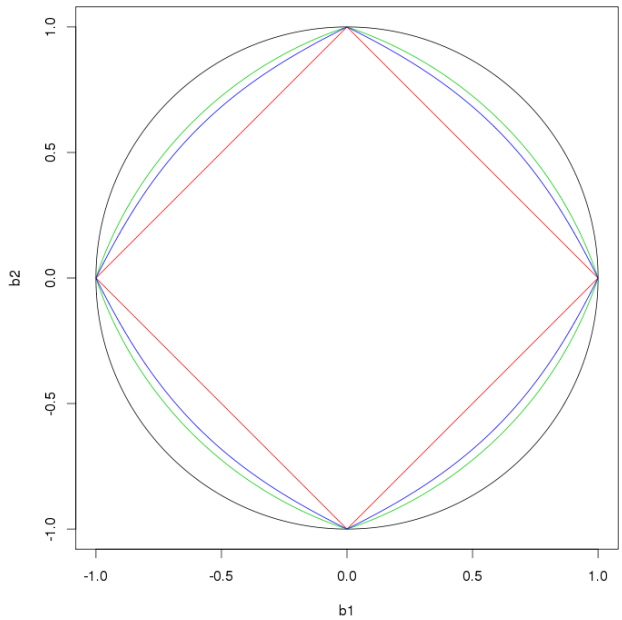

Figure 6: Contour of solutions for penalty=1, ridge (black), lasso (red), elastic net ($a = 0.5$), (green), PED (blue).

From Fig. 6, we can see how the elastic net penalty and the PED's geometric mean norm lie between the lasso and ridge contours. The elastic net penalty will vary depending on the ratio of the two penalties employed. For the 2 dimensional example given the PED and elastic net contours look very similar

and there will be a mixing level of the elastic net that is very close to the PED penalty, however this is not generally the case.

Fig. 7 represents the cross section of the penalties along the line $b_2 = 0$. We plot the penalty level for $b_1 \in [-1.5, 1.5]$ to show how the different penalties behave along the axes when the covariates grow. We can see that along the axes, the PED penalty equates to the ℓ_1 norm. i.e. the function is linear along the axes. Hence, when the problem is very sparse and there is only one non-zero parameter, the PED obtains the square root lasso estimate.

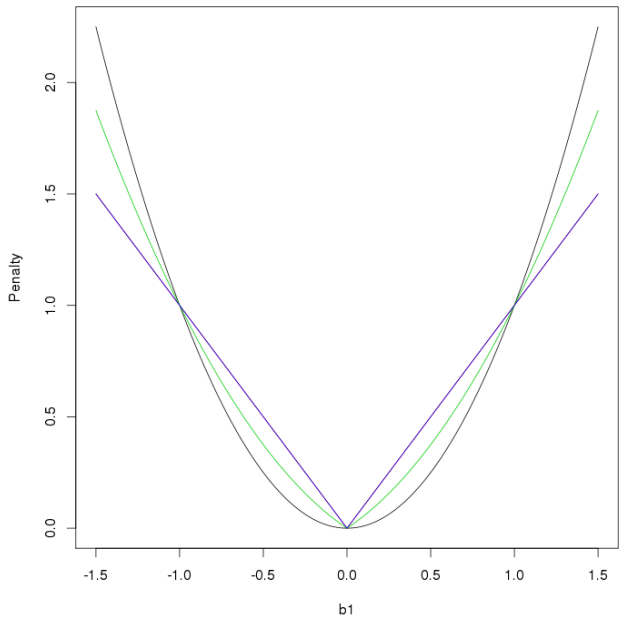

Figure 7: Cross section along axis for varied b_1, $b_2 = 0$ fixed. Ridge (black), elastic net ($a = 0.5$), (green), PED/ lasso (blue).

The penalties intersect at 1, but the ridge (and naturally the elastic net) penalty is quadratic along the axes. Again the elastic net becomes closer to the lasso/ridge penalty depending on the level of mixing. We can also see that for a penalty level of greater than 1, the ridge penalty becomes more restrictive on the size the estimated parameters can take. In contrast, the penalties involving

the ℓ_1 norm are less restrictive around the axes (i.e. when models are sparse). This can be seen more clearly when we consider the contour plots for larger values of the penalty (Fig. 8).

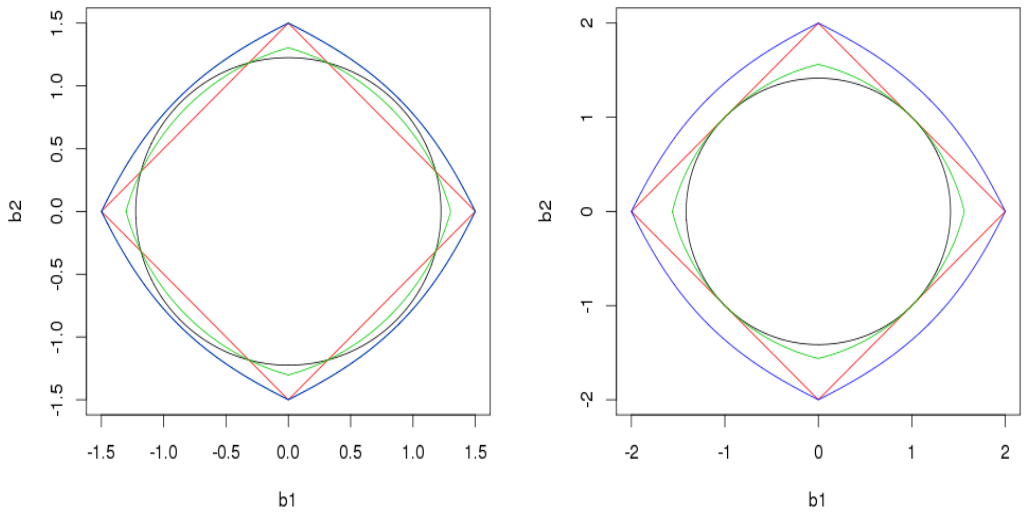

Figure 8: Contour plots: penalty=1.5, penalty=2, ridge (black), lasso (red), elastic net ($a = 0.5$), (green), PED (blue).

2.2 CHALLENGES IN MEG

One of the more established source localisation methods used for MEG data is the minimum norm which in its unconstrained form reduces to ridge regression. Of course under this minimum norm setting, every location will have a non-zero signal estimate, whilst the shrinkage property means that for any source that is located, the activity will be spread across a number of spatially correlated locations— something known as 'signal leakage' (Wens et al., 2015). Therefore, given the basis of the minimum norm in ridge regression, sparse regression methods provide a natural extension to the existing methods.

It is worth spending some time to give more detail to MEG data and the types of activity that we are often trying to identify. In MEG data analysis we are typically trying to detect small electromagnetic signals in the brain against a background of potentially significantly higher noise levels (Vrba, 2002). Whilst some of the noisier artefacts of the data can be accounted for by employing magnetic shielding and factoring for the underlying magnetic rhythms of the body, we are still left with data that is considerably noisy. One of the ways to reduce the noise level of the data is to average over trials. Assuming the noise is not temporally correlated between trials, averaging increases the signal-to-noise ratio by $\sqrt{N}$ when N is the number of trials (Gonzalez-Moreno et al., 2014). However, averaging is not always the best approach to take with MEG data.

Generally there are two different types of signals that can be detected in MEG data. When we average data we are assuming there is something that resembles an average response that we will be able to detect. This is the case for evoked responses, where under a particular stimulus the observed response will typically be similar each time. Such responses are relatively simple to time lock so we can be confident that the activity will occur in roughly the same area temporally. This means that averaging trials with such a response will reduce the noise in the data whilst maintaining the important spikes in activity. However, this approach becomes less appropriate for data where the time or phase of the response is not locked over time. Induced responses characterise activity which is temporally connected to a stimulus, but where the phase information is not fixed (David et al., 2006). In oscillatory signals (a typical form of induced response) if the max/min occur at slightly different times over trials then averaging may cancel out any increase in signal. The focus for induced response data is thus on the relationship between the power and time. There-

fore amplitude sensitive and time-frequency analyses are frequently employed for induced response data (Litvak et al., 2013).

Given the different types of data that can be observed, it is worth noting that certain methods are more appropriate to certain types of data. Beamforming relies on the covariance information of the data and the resulting variance change between areas of activity and inactivity. However, the accuracy of the beamformer estimate is dependent on the length of the covariance window. Therefore, since averaging naturally reduces the number of time points in the data and removes any induced activity it actually inhibits the beamformer (Brookes et al., 2010). For this reason, beamformer methods for evoked responses tend to compute the covariance matrix over all trials in order to construct the estimate (Robinson, 2004; Cheyne et al., 2007). These properties mean that beamforming is quite computationally expensive for evoked response data and it is more suited to induced activity. On the other hand with the sparse regression methods we will essentially treat each time point as an independent observation, therefore averaging over trials for evoked responses seems the most appropriate area for their application.

Of course one of the disadvantages of this type of data is the need to record multiple trials. As we have already noted, the larger the number of trials we use the greater the SNR becomes, however ideally we would also like to keep the number trials small in order to reduce the time and cost of experiments. Therefore, we would like to find a method that is robust to the larger noise found when using a smaller number of trials. With this in mind, one interesting aspect to investigate is how the estimates for the implemented methods change as we reduce the number of trials over which we average.

2.3 SIMULATIONS

We base our simulation on a typical MEG neuroimaging experiment. We consider a single slice of the head where each location is a potential source that needs to be estimated and, to begin with, a single source is simulated over 500 time points (the second 250 time points were a negative version of the first 250 in order to have the zero mean over time that is necessary for beamforming). We then add additional sources nearby and adjust the level of noise involved. In order to assess the accuracy of the different methods we want to look at the number of true positives and the number of false positives. The true positives represent the number of locations of true activity that have non-zero estimates and false positives are the number of non-zero estimates that are zero in the true model. For these sparse models we want a small number of false positives. Additionally we look at the root mean square error (RMSE) as it includes both variance and bias into its measure.

In the following simulation, we use a single slice from a typical head dataset to represent a section of the head that we are interested in. This slice consisted of 278 locations of interest— each a 8mm by 8mm voxel. At each of these locations we have a leadfield matrix of dimensions 270×2 which gives the forward solution for 270 sensors from the contributions of the location's activity in two orthogonal orientations. By using two orientations we would be able to reconstruct both the source strength and direction information. Therefore, we have 556 parameters that we need to construct.

2.3.1 *Single source simulations*

A single source was simulated with strength 5. Then random noise was added with variance $\sigma^2 = 2$ for 500 time points. The minimum norm (ridge), elastic net and lasso were fitted via the glmnet R package (Friedman et al., 2010) with the penalty value chosen by cross validation.

The square root lasso used the asymptotic value of λ and was implemented using the slim function in R package flare (Li et al., 2014). The asymptotic penalty, $\lambda = 1.1n^{1/2}\Phi^{-1}\left(1 - \frac{\alpha}{2p}\right)$, is chosen to give near-oracle performance in the square root lasso and, due to the pivotal nature of the square root lasso, the penalty is independent of the noise variance σ^2 (Belloni et al., 2011).

Two methods of penalty choice are included for the PED; BIC and the ξ method of maximising $\hat{k} = \sqrt{\frac{||\hat{\boldsymbol{\beta}}_{PED}||_2}{||\hat{\boldsymbol{\beta}}_{PED}||_1}}$. The value of the thresholding constant C and the penalty level λ in the PED method was determined by a grid search (using $C \in \{0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.2\}$ and 50 values of λ equally spaced between 0.001 and 1.5), where the chosen pair C, λ corresponded to the smallest value of BIC or the largest value of $\hat{k}$ respectively. This combination of threshold and penalty level was then used across the data to determine when a parameter should be set to zero. To save computational time, we only looked at the first 50 time points. Fig. 9 shows the true source distribution and the means of each estimate across the first 50 time points. We also include the beamformer and the modulus of the beamformer estimate.

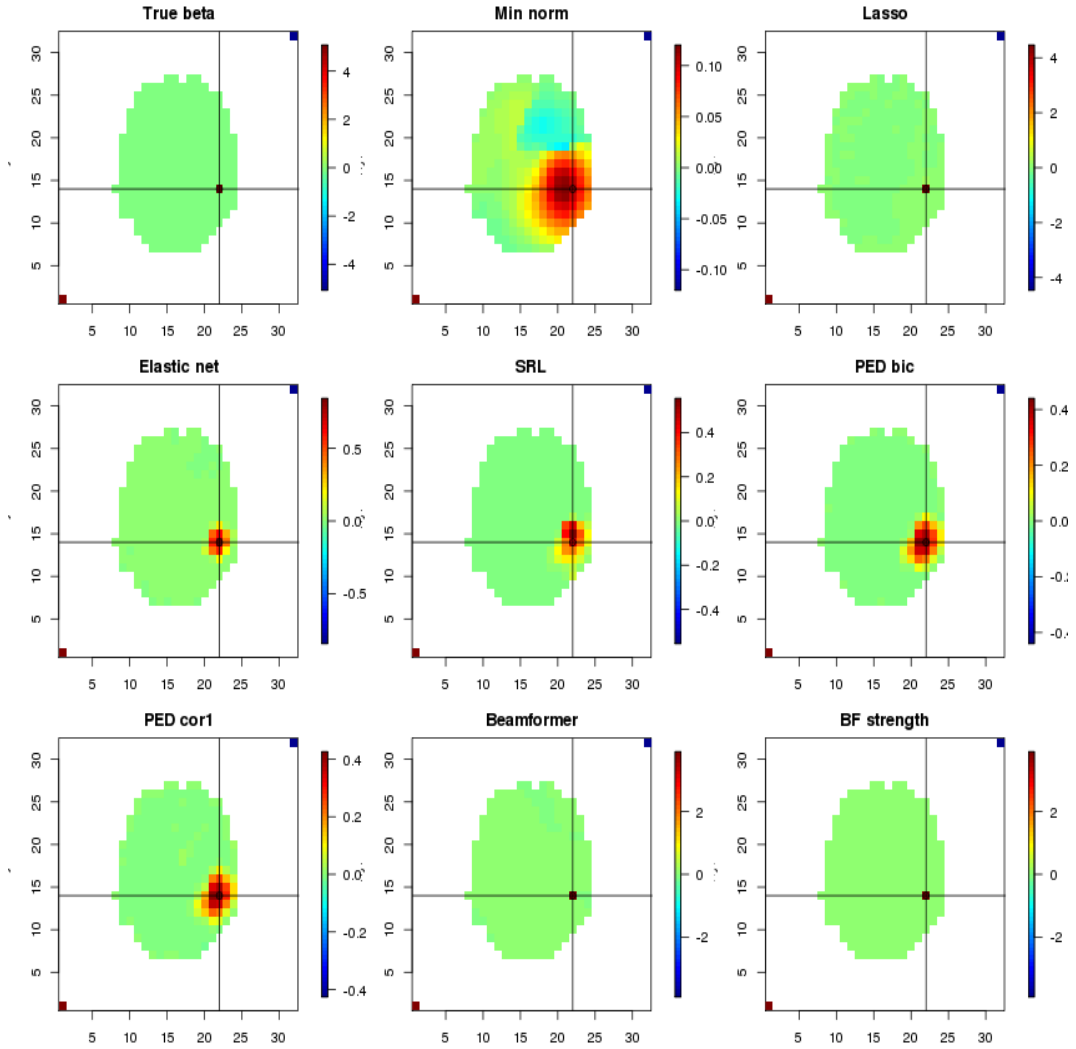

Figure 9: Means for single source, $\sigma^2 = 2$

Table 1 contains the mean number of false positives for each method across the time points as well as the root mean square error. In Table 1 the RMSE was calculated by

$$RMSE(\hat{\boldsymbol{\beta}}) = \frac{1}{50}\sum_{t=1}^{50} ||\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_t||_2,$$

where $\boldsymbol{\beta}_0$ is the true source vector (note this was unchanged over the time points) and $\hat{\boldsymbol{\beta}}_t$ is the vector of source estimates for time point t. The mean square error is a useful measure of how accurate an estimate is as it combines both the bias and variance of an estimate. However, it may not be as suitable for the beamformer, which bases its estimation on the variance of the data.

In Table 1, despite the relatively accurate mean, the beamformer RMSE is the highest due to its variability. This could partly be attributed to the short covariance window used.

Method	Mean TP's	Mean FP's	Sum RMSE
MN	1	277.00	11.80
Lasso	1	11.26	2.20
EN	1	20.8	9.15
SRL	1	28.38	9.43
PED-BIC	1	30.66	10.07
PED-ξ	1	36.38	10.73
BF	1	277.00	12.93

Table 1: Summary— True and False positives, and RMSE for single source.

In the single source example the lasso performs very well in both the number of false positives and RMSE. Its deficiencies in terms of the treatment of correlated variables are less relevant in this case as it was only required to identify a single source. However we would expect issues if we wanted to identify multiple sources from spatially correlated locations. The other sparse methods produce more spread out estimates, however all of the sparse methods show an improvement over the ridge estimate. In terms of accuracy, there is very little separating the elastic net, square root lasso and penalised Euclidean distance. The primary regions of activity are roughly equivalent across these methods and with the exception of the SRL (which places the peak one pixel away) the peak is consistently placed in the correct location. The number of false positive locations of activity is perhaps a little high, with the ξ selection method of

PED performing the worst, however as we can see from the plots, the majority of these false positives are also very close to zero.

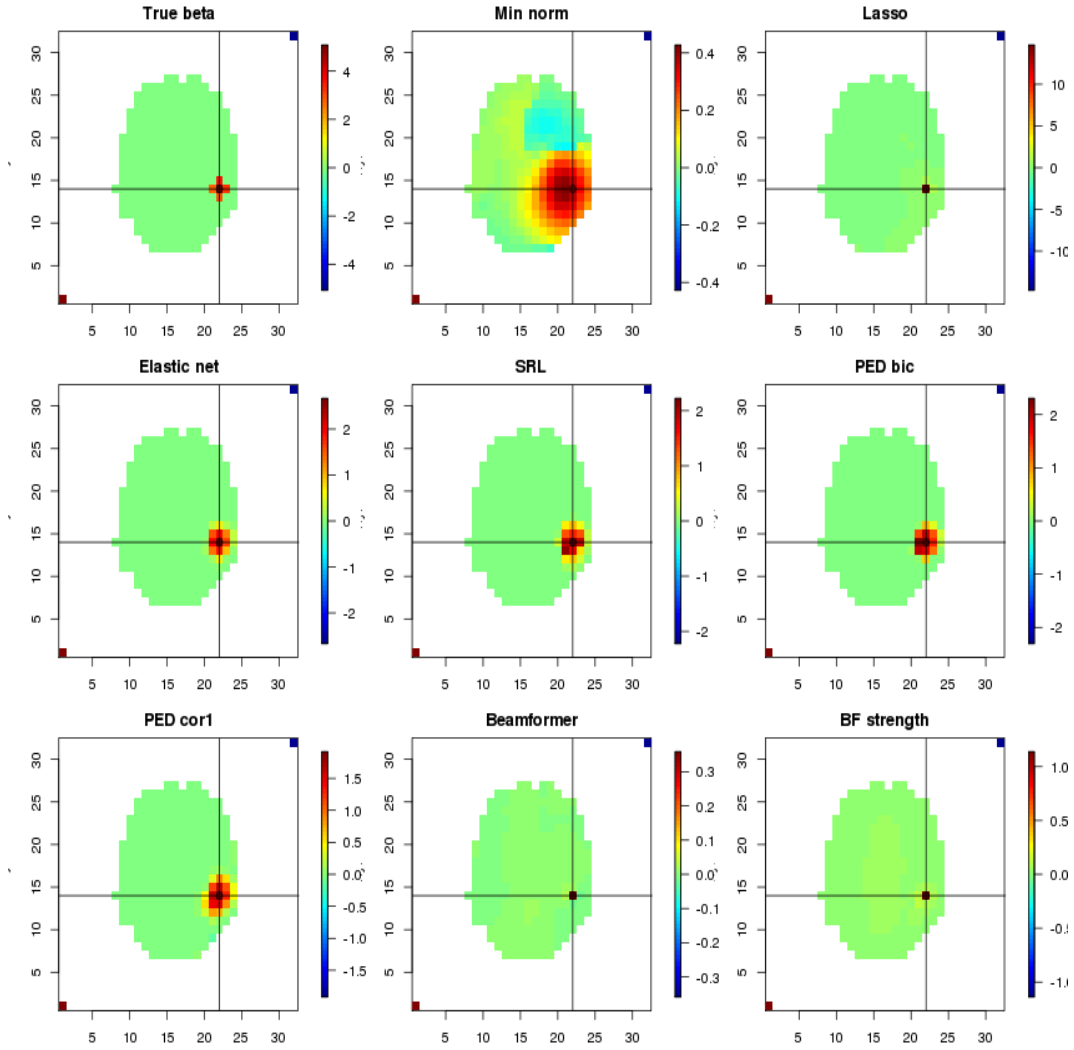

Figure 10: Means for group of sources, $\sigma^2 = 2$

2.3.2 *Group of close sources*

In the same way as the previous simulation, we now simulate a group of sources. The methods are implemented as before. Additionally, we perform the simulation under different levels of noise; $\sigma^2 = 2, 5, 15, 100, 400$. Figs. 10-11 show the mean estimates for a number of these noise levels.

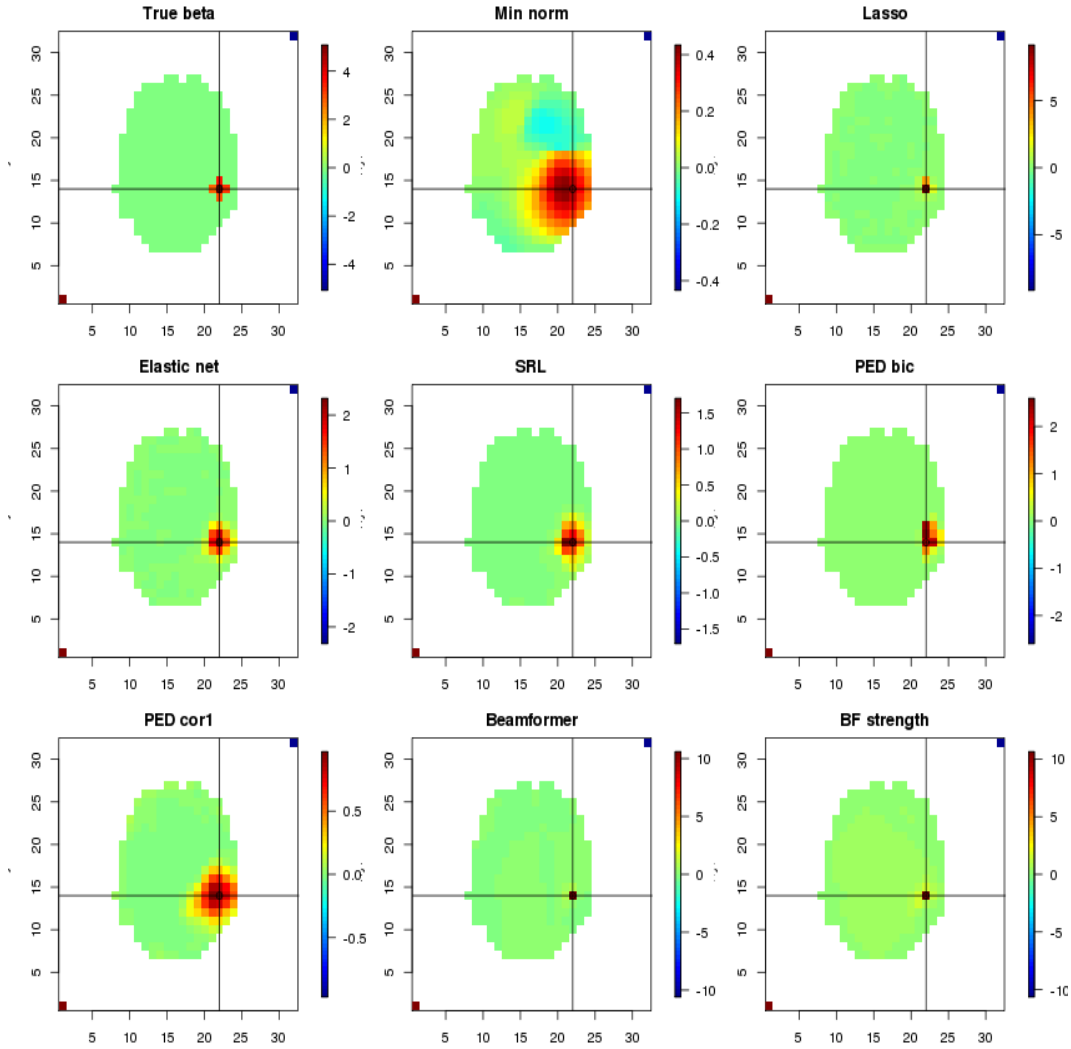

Figure 11: Means for group of sources, $\sigma^2 = 400$

Noise		MN	Lasso	EN	SRL	PEDb	PEDξ	BF
	TP	-	4.86	5	5	5	5	-
$\sigma^2 = 2$	FP	-	10.5	13	16.1	9.26	22.02	-
	RMSE	39.03	20.22	15.71	18.56	16.93	21.28	27.42
	TP	-	4.62	5	5	5	5	-
$\sigma^2 = 5$	FP	-	9.94	13	17.38	9.98	28.32	-
	RMSE	39.04	21.12	15.90	19.21	18.64	22.92	35.07
	TP	-	4.08	5	5	4.78	5	-
$\sigma^2 = 15$	FP	-	11.66	15	19.86	9.80	33.18	-
	RMSE	39.09	21.66	16.80	19.84	14.43	25.58	43.82
	TP	-	3.78	5	5	4	5	-
$\sigma^2 = 100$	FP	-	12.22	22	22.58	10.48	43.04	-
	RMSE	39.54	26.13	21.93	21.42	15.11	30.67	89.37
	TP	-	3.34	5	5	4.06	5	-
$\sigma^2 = 400$	FP	-	9.32	23	24.52	10.04	55.84	-
	RMSE	41.21	34.32	28.80	22.92	22.44	38.93	157.41

Table 2: Summary-Mean true positive locations, mean false positive locations and sum(RMSE) for group of sources under various noise levels.

As we expect, the lasso gives very sparse solutions and tends to attribute the source strength to one location. Although it does seem to have chosen most of the correct locations, as evidenced by the number of true positives, the absence of a grouping property in the lasso means that almost all of the activity from the cluster of sources is assigned to a single location. The PED under BIC and the elastic net seem to perform the best in terms of sparsity and accuracy over the different noise levels. The PED (BIC) method is the most resilient to the changes in noise variance as both the RMSE and the number of false positives remain fairly consistent. The elastic net, PED (BIC) and the square root lasso seem to give the best performance in terms of RMSE, with the PED estimate being particularly consistent in this regard.

Note the beamformer seems to have performed poorly, although the scale of the estimates improves as we increase the noise variance. The reason for the first point is that the beamformer is noted to struggle with highly temporally-correlated sources and looks for a single location for which to attribute the source. Therefore it will find the 5 perfectly correlated sources in this example challenging. Secondly, as a covariance based spatial filter, the beamformer requires a certain level of variance in the data in order to construct the estimates. This is confounded by having a short window over which the covariance is estimated. For a similar reason, we tend not to pre-average the data over trials for the beamformer as it can mask certain features when it comes to producing the beamformer estimates.

2.3.3 *Two distant sources*

The beamformer's issues with time correlated sources become more evident when we consider more spatially defined sources. In the following simulation two perfectly correlated sources were simulated; one in the left and one in the right of the brain. The sources were simulated with a strength of 5 on a single slice of the brain for 2000 time points and random noise was added to the data with a standard deviation of 10. The mean estimates for the minimum norm, beamformer and sparse regression methods are found in Fig. 12, with the corresponding variances in Fig.13.

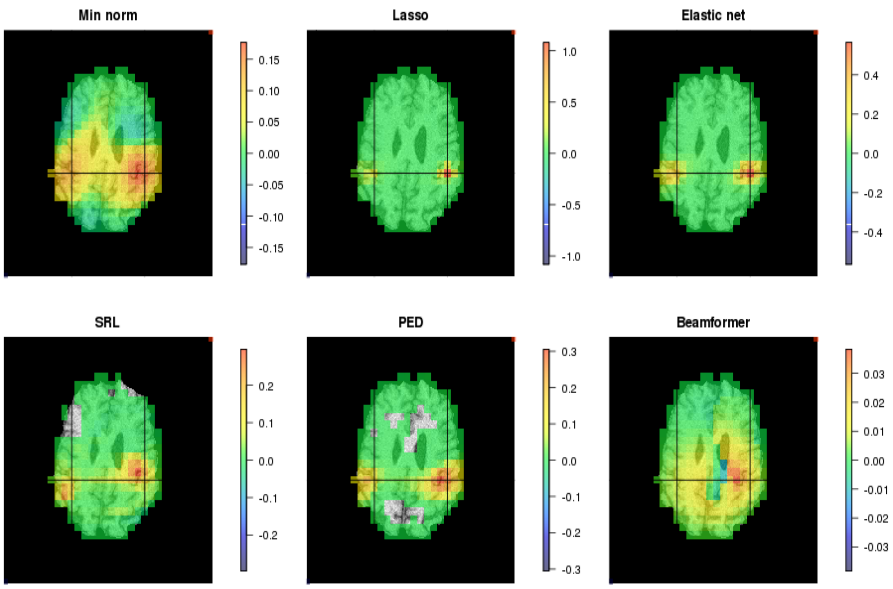

Figure 12: Mean estimates of two correlated sources over 2000 time points. Cross-hairs denote locations of true sources.

The mean estimates for the sparse regression methods generally perform well in identifying the correct areas of activity, although interestingly they seem to place a stronger right source. The square root lasso's placement is a little erroneous for both sources and the minimum norm displays some characteristic

leakage between the sources, but each of the regression methods has clearly identified two sources. The failure of the beamformer in these circumstances is evident as it places a single source slightly right of centre. Furthermore, the strength of the activity is extremely low, even compared to the very spread out minimum norm.

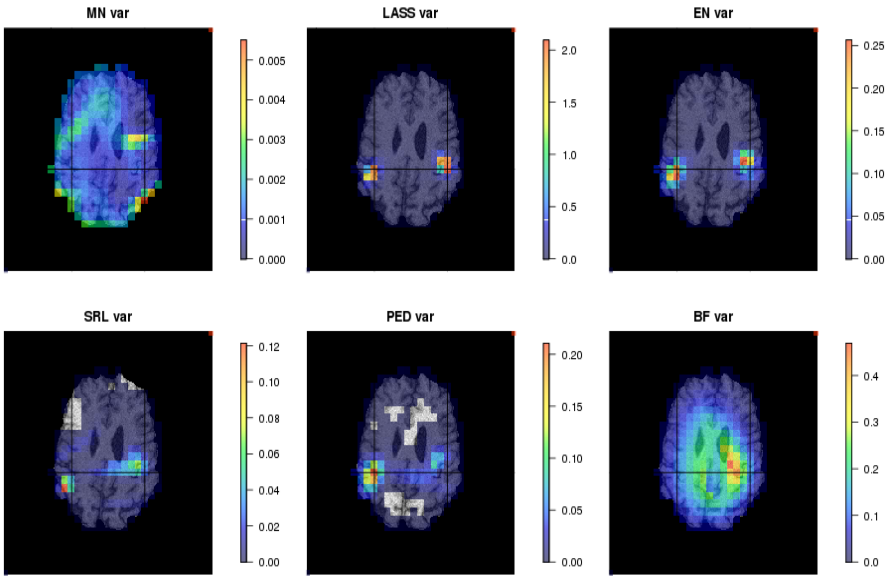

Figure 13: Variance of estimates of two correlated sources over 2000 time points. Cross-hairs denote locations of true sources. Note on legend scales: since the variances from the sparse MEG estimates vary so widely, I have chosen not to apply the same scale across MEG variance plots for different methods as to do so may hide features of how the variance is distributed according to the location.

The variance plots show that despite the mean estimate being low in the beamformer, the variance is comparatively high around the location of the placed source. Nevertheless, the largest areas of variability in the beamformer estimates are mainly found around the central areas of the slice rather than the true source locations. In contrast, the sparse methods have regions of higher

variance that correspond to their placed sources, which are much closer to the true locations. Therefore, sparse regression methods perform much better than the beamformer when there are highly temporally-correlated sources, with the lasso and elastic net in particular giving good results.

2.4 REAL DATA

To test our methods with some real data we have a dataset of median nerve stimulation evoked response recordings. We have 1132 trials each of time 0.5 seconds with a sample rate of 2400 Hz. In each trial the left median nerve was stimulated via a small electric shock. So we have a trial length of 1200 time points and 1132 trials. Each of the methods are implemented for the full number of trials over suitable sections of the brain. They are then considered again using reduced number of trials. Fig. 14 shows the trial averaged data from the 270 MEG sensors around the head.

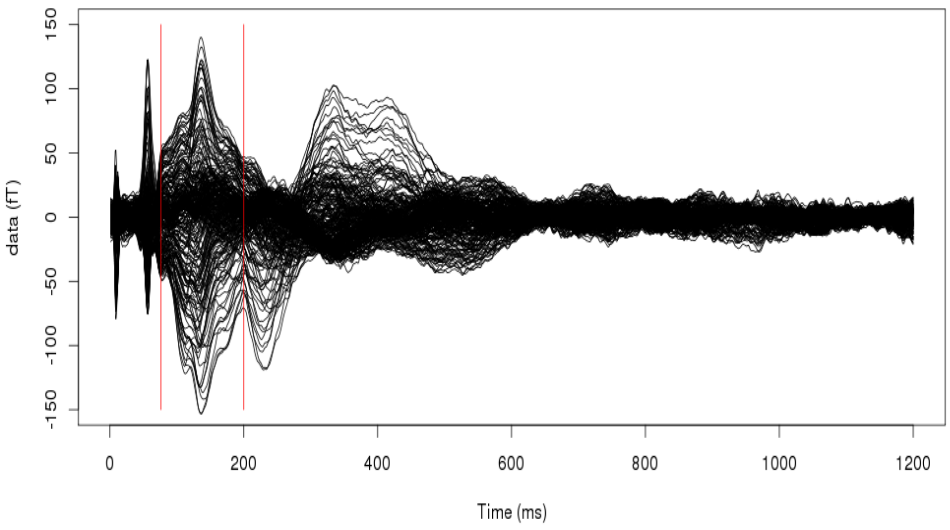

Figure 14: Data recordings from all 270 sensors when averaged over 1132 trials.

The assessment of the model estimates in the real data case is more problematic than in the simulations as in general we are less confident of where the sources are. However, the cortical response to this type of stimulus is well studied and we would expect the activity to be centred around the left hand somatosensory cortex on the postcentral gyrus (roughly the centre right edge of our slices) (Waziri et al., 2005). However, from a statistical point of view there are no established measures of accuracy for the fit of the models under this real data. In terms of assessing the change in the models over varying number of trials we can examine the change in mean estimate compared to the fit under the full number of trials.

2.4.1 *Single slice*

To begin with a single slice in the head was chosen to compare our methods. For all the penalised methods the data was averaged over all 1132 trials and the estimates for each method were then computed for the primary window of activity (window between red vertical lines in Fig. 14). Subsequently, this process was repeated for a reduced number of trials (1000, 500, 250, 100, 50, 25, 10, and 5) in order to assess how the methods changed under fewer trials (and thereby greater noise). All analysis was performed in *R*. We included two methods of choosing the sparsity parameter for PED (BIC and ξ), the elastic net, minimum norm and lasso penalties were chosen according to cross validation using the *glmnet* package and the square root lasso was fitted with the *slim* function using its asymptotic lambda value as noted in section 2.1.5. Fig. 15 presents the mean estimate over the 125 time point window for reducing number of trials using the PED (BIC) method.

The principal area of activity in the first image is roughly where we would expect to see signal in this dataset, however, it gradually changes as we reduce the number of trials. By about 50 trials, a new area of activity is present in the estimate and for less than 25 trials the estimate has changed considerably.

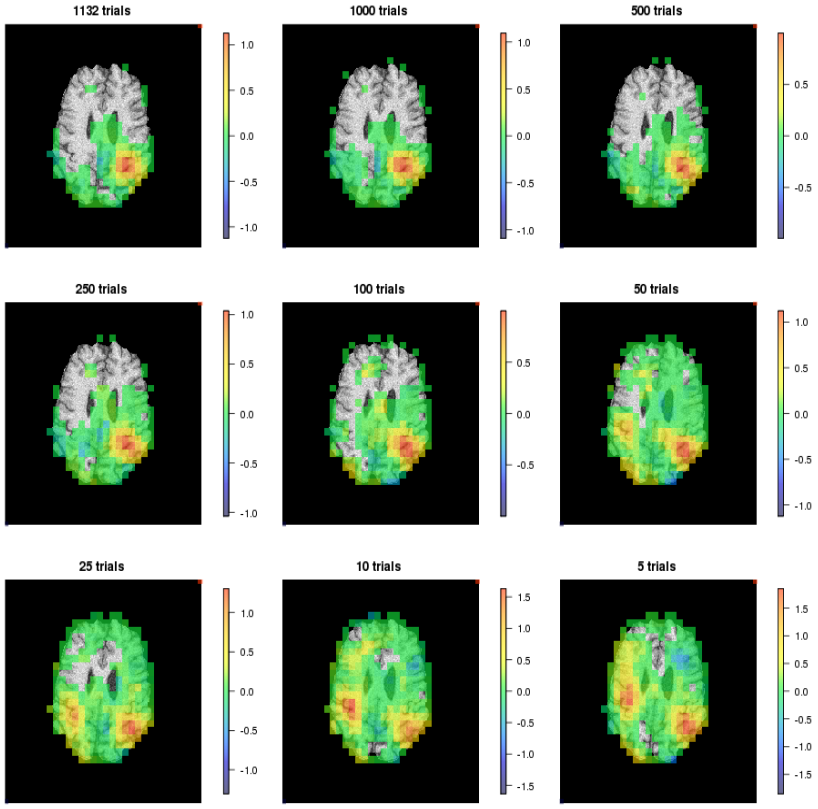

Figure 15: PED (BIC) estimate for 1132, 1000, 500, 250, 100, 50, 25, 10 and 5 trials respectively (left to right, top to bottom).

The most consistent method was the lasso, which remained largely similar in its placement of activity even at 25 trials. However even this estimate begins to introduce more spurious sources when we get as low as 10 trials. In order to summarise the change in the estimate when the number of trials was changed, we look at the mean difference between the estimates for the full and reduced trials. Fig. 16 shows plots of the mean difference for each method normalised using the strength of signal estimated under the full number of trials. i.e. if

$\hat{\boldsymbol{\beta}}_f$, $\hat{\boldsymbol{\beta}}_r$ are the estimates using the full and reduced number of trials respectively, in Fig. 16 we plot $\frac{||\hat{\boldsymbol{\beta}}_f - \hat{\boldsymbol{\beta}}_r||_1}{||\hat{\boldsymbol{\beta}}_f||_1}$ for each method and number of trials.

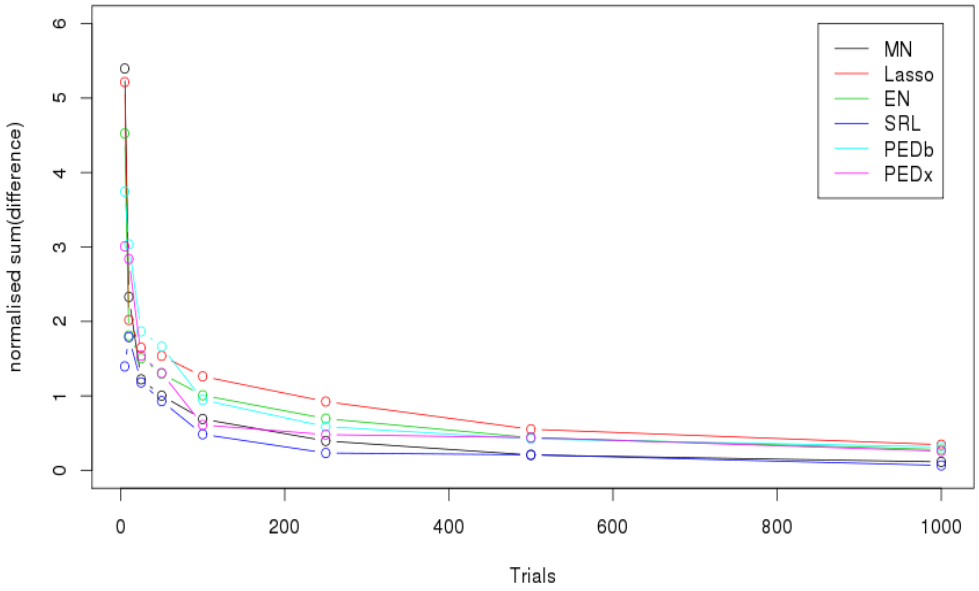

Figure 16: Normalised mean difference between 1132 trials and 1000, 500, 250, 100, 50, 25, 10, and 5 trials.

We can see that the square root lasso is the most consistent over fewer trials in terms of mean difference. The lasso seems to be the worst performing in this regard, despite being very consistent in the general area of activity placed. This is due to the peak of the lasso jumping between a clump of locations, so whilst the general area remains consistent, the exact pixel of the peak changes from the original estimate. What is evident is that the penalised regression methods all give fairly consistent estimates down to around 100 trials, and that for less than 50 trials the estimates begin to diverge considerably. We now extend the implementation of the sparse methods to cover a larger area of the brain, thereby increasing the number of parameters.

2.4.2 *Larger scale analysis*

We up-scaled the number of locations of interest to include a much larger section of the brain. The estimates were now computed over 5 brain slices totalling 1367 locations of interest and 2734 parameters. Again the activity was expected in the mid/low right hand corner of the images, however since our area of interest was now 3-dimensional, the lowest slice was considered to be too deep to be the source of activity. The estimates were computed for the entire 1200 time point recording using all 1132 trials. In all methods with the exception of the beamformer, where the data covariance over the entire data recording is required in the computation, the data was averaged over trials beforehand. The beamformer estimate was weighted according to its depth/location by dividing the estimated time course at each location θ by $\sqrt{\mathbf{w}_\theta^T \mathbf{\Sigma} \mathbf{w}_\theta}$, where $\mathbf{w}_\theta$ is the corresponding weight for the location as defined in the beamformer theory (section 1.6, equation 1.6) and $\mathbf{\Sigma}$ is the diagonal estimated noise covariance matrix. Figs. 17-18 present the mean estimates for each method over the primary active period (time points 76-200).

From the images below, we can see that the placement of sources in the active period seems to be consistent across all the methods, however the distinction between the methods comes in the distribution of the activity. The lasso is considerably more sparse than the other methods in its distribution of the activity. This can be seen in the very small sharp peaks of activity. At the other end of the scale, the minimum norm spreads out the activity across a very large area.

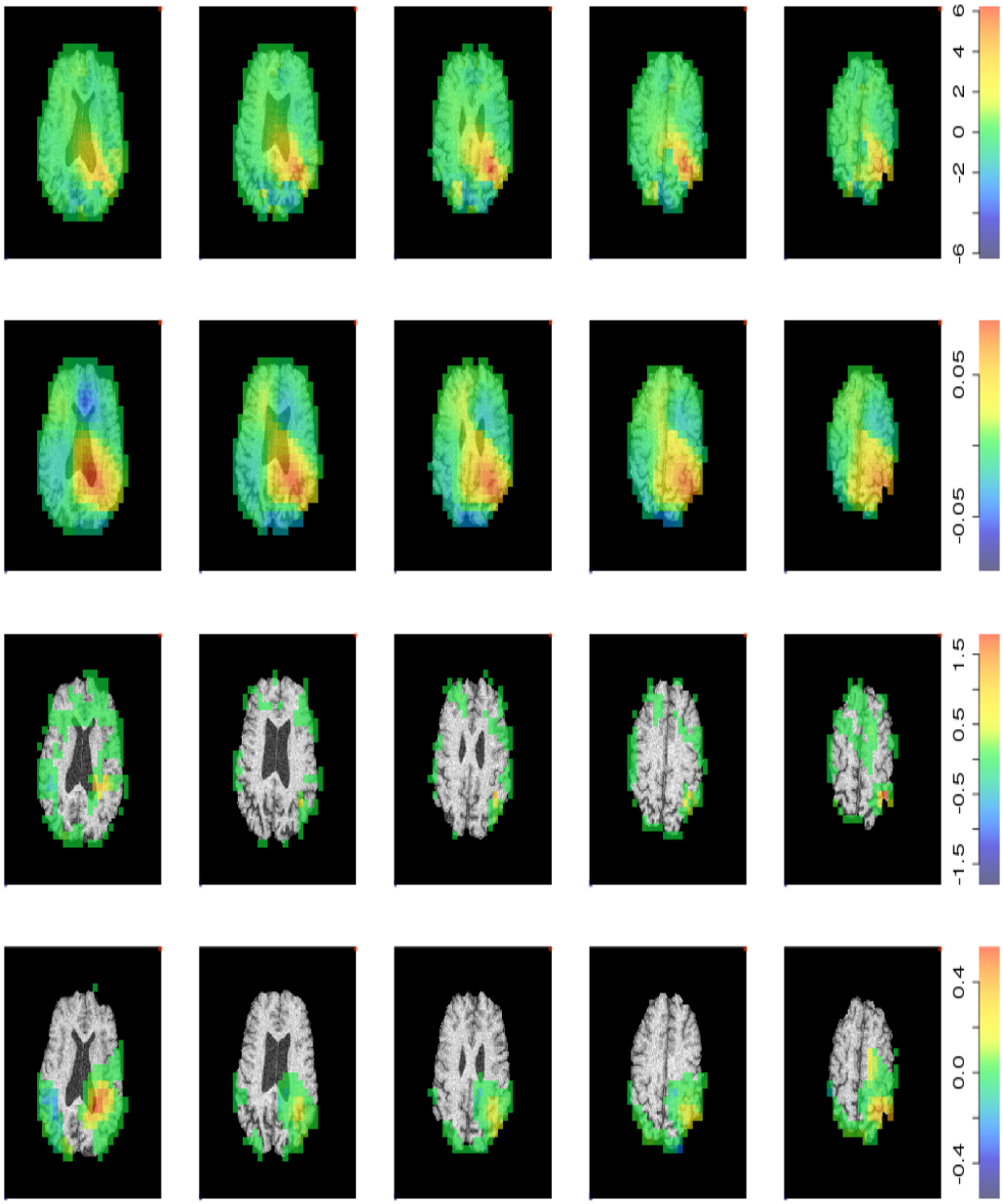

Figure 17: Means for (top-bottom) beamformer, minimum norm, lasso and PED (BIC)— Active window (pts 76-200). Columns from L-R lowest slice to highest slice in head.

The elastic net, PED and SRL methods produce estimates that are more spread out than the lasso but much more localised than the minimum norm and even the beamformer.

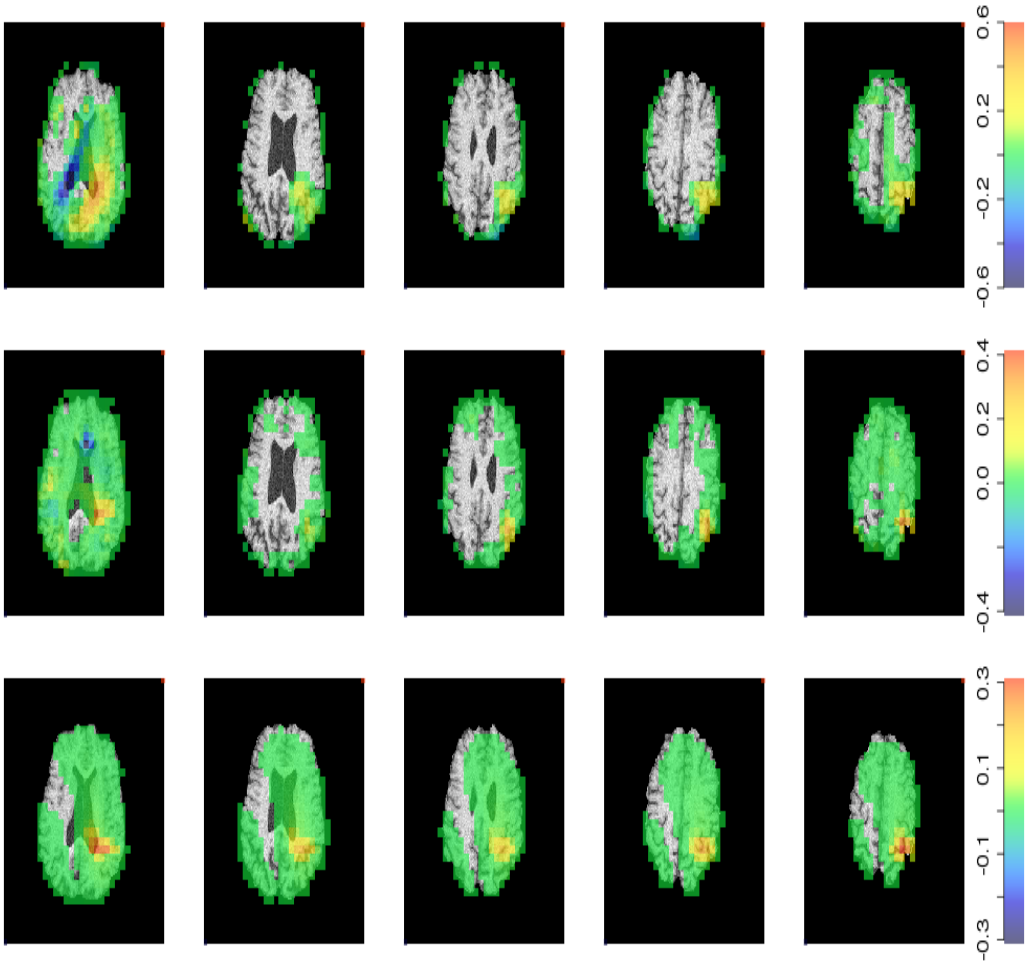

Figure 18: Means for (top-bottom) PED (ξ), elastic net and square root lasso—Active window (pts 76-200). Columns from L-R lowest slice to highest slice in head.

Less encouraging however, is that all the methods seem to place sources of activity in the lowest slice (first column in Figs. 17-18). This is particularly evident in the PED, minimum norm and square root lasso methods. This section of the brain is too deep to expect activity directly related to our stimulus. Therefore the activity either has another source, or there is bias present. Comparing the active and the control windows, it seems that the estimates in the control window (where there is no direct stimulus active) are much larger for

the lowest slice than the others across all methods (save the depth weighted beamformer).

Method	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5
BF*	22.1	22.6	22.0	19.2	14
PED-BIC	80.7	11.9	2.2	2.7	2.5
PED-ξ	88.1	3.2	2.3	1.4	4.9
SRL	92.6	4.0	1.4	0.7	1.3
Lasso	94.6	2.2	0.7	0.6	1.9
EN	90.2	3.4	2.1	1.5	2.8
MN	70.9	16.3	5.8	4.2	2.8

Table 3: Percentage variance covered by slice (1 is lowest, 5 highest) for control window (time points 901-1200). *Note beamformer has had depth weighting applied. The number of locations in each slice is as follows (322, 314, 278, 249, 204).

In fact the higher four slices display virtually no activity at all. Furthermore, the variance over the control window for the lowest slice is roughly 4 times the variance in the other slices (see Table 3). This suggests that at least some of the activity in the deeper sections presented may be partly due to other underlying activity (or noise) in combination with the true activity from shallower locations. Therefore, this is most likely a feature of the data rather than an error in the estimation across the methods. The variance over the control window could be used to inform a depth weight correction to be applied over the estimates. We will give some thoughts to how this could be done in the discussion section at the end of this chapter.

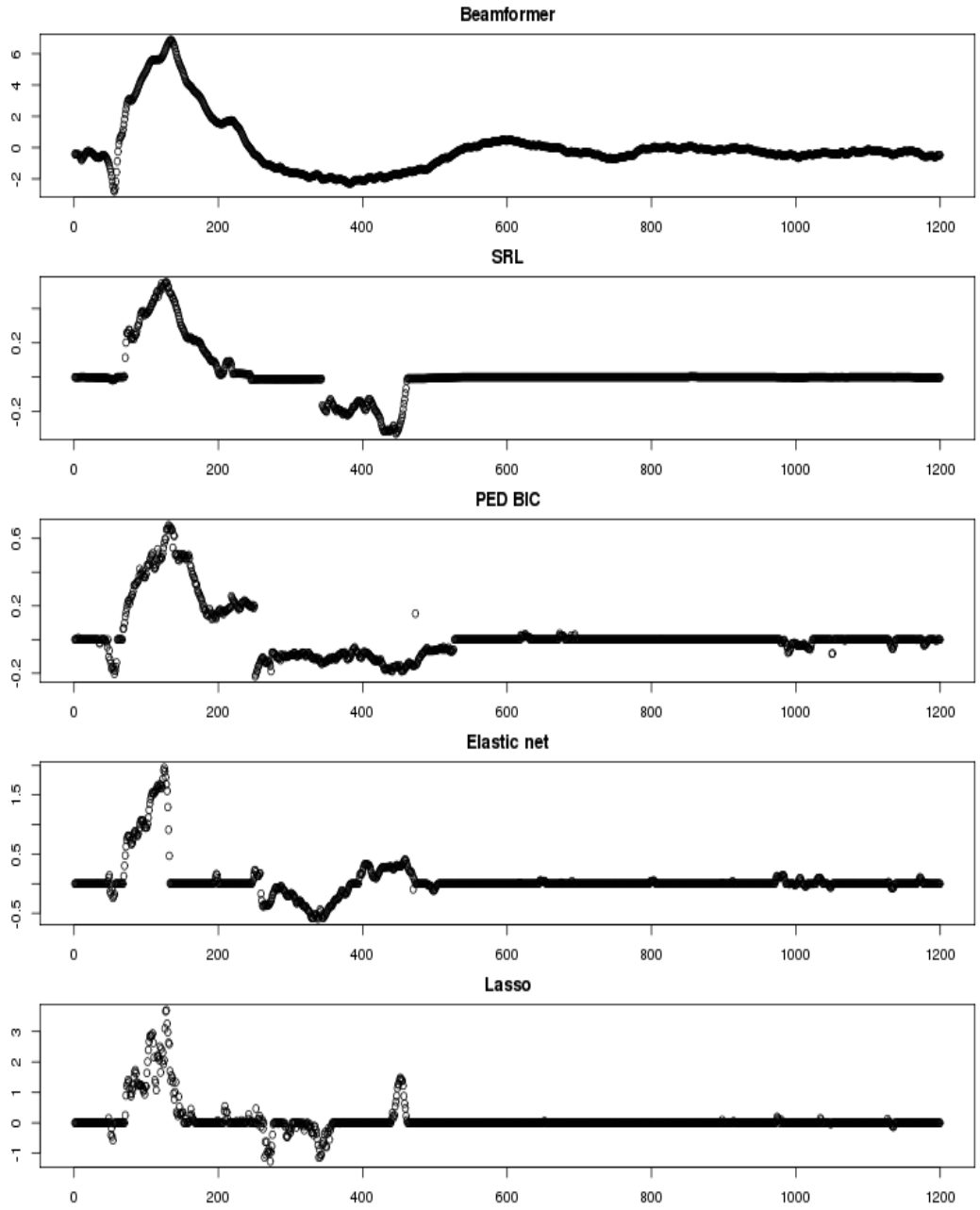

Figure 19: Time courses for beamformer, square root lasso, PED (BIC), elastic net and lasso.

Another thing to consider is the temporal aspect of the sparse methods. Fig. 19 shows the reconstructed time courses at location (20, 12) in the uppermost slice for the beamformer along with the square root lasso, PED, elastic net and lasso sparse methods.

As we can see the general shapes of the time courses are consistent across the methods, however it is evident that the sparse methods suffer from a lack of smoothness in some parts and the curves are relatively spiky (this is particularly the case in the lasso). This is naturally a downside of sparsity in that when we treat each time point in isolation the selection of covariates will vary, even for neighbouring time points. In the PED example, there is also an evident sign change in the time course that is due to the sign changing in one of the components of the source.

2.5 DISCUSSION

Sparse methods provide an alternative approach to the minimum norm that reduces the dimensions of the problem by returning parameter estimates that are equal to zero. The minimum norm estimates tend to be much too smooth for giving precise source localisation. The sparse methods all provide an improvement in estimation over the ridge (minimum norm) method in MEG data. In the simulations, under various noise levels, the methods are generally quite successful at locating the true location of the sources. The lasso performs the best for very sparse examples, but fails to accurately reflect the topography of the signal when we have multiple highly correlated sources. For these examples the elastic net, square root lasso and penalised Euclidean distance are more suitable. The PED under BIC performs especially well in terms of both number of false positives and RMSE. These results support the findings in the original paper (Vasiliu et al., 2014) where, when compared to the elastic net and lasso in simulations, the PED had fewer false positives. However the lasso and EN occasionally perform better in terms of true positives.

The sparse methods have generally attributed activity to sensible areas, but the estimates are not without their problems. As well as locating activity in the areas we would expect, all the methods tend to place a strong source of activity much deeper towards the centre of the head. This issue is not reserved to these sparse methods as the minimum norm displays similar behaviour. This is somewhat surprising since the standard minimum norm estimate has a well documented issue with placing superficial sources (Lin et al., 2006). Additionally, Uutela et al. (1999) suggests that this issue may also be present in ℓ_1 norm situations. This suggests that our initial assumptions about the noise of the data may be simplistic. These assumptions fall down when we consider the possibility that the noise is also dependent on either the sensor location, or the variability of the sources themselves through, say depth in the head. The larger variance in the deeper slices demonstrated in the analysis above seems to lend some weight to this hypothesis.

One potential way of addressing this issue is to employ some sort of depth weighting in order to normalise the estimates. This is frequently employed in beamforming by normalising the estimate using the relevant weight vectors and the estimated noise covariance in order to account for greater variability in deeper areas of the brain. When this was employed in the real data analysis earlier, it seemed to go some way to reducing the variance originating from deeper locations, as the variance estimates were comparable across all slices. Furthermore, the deeper activity appears to be much less prominent in the weighted beamformer than the other methods. The approaches to weighting in the beamformer and minimum norm are outlined in chapter 1. Since the solutions for the minimum norm and beamformer are available analytically, the weighted estimates are usually obtained by dividing the existing solution by

some weight. For the sparse methods, the estimate is obtained by numerical optimisation, therefore producing a weighted estimate is much more problematic. For the weights normalised beamformer, the weights were chosen to give unbiased noise variance estimates (Luckhoo et al., 2014) and it seems reasonable that the same objective would be applied to a weighted estimate for the sparse methods. Therefore, for a weight matrix $\mathbf{W}$, taking

$$\hat{\mathbf{s}} = (\mathbf{W}^T\mathbf{W})^{-1/2}\mathbf{W}^T\mathbf{d},$$

gives

$$\text{Var}(\hat{\mathbf{s}}) = \sigma^2(\mathbf{W}^T\mathbf{W})^{-1/2}\mathbf{W}^T\mathbf{I}\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1/2} = \sigma^2\mathbf{I}.$$

In the objective function for the sparse regression method we then replace $\hat{\mathbf{s}}$ with $\frac{\hat{\mathbf{W}}^T}{||\hat{\mathbf{W}}||_2}\mathbf{d}$. i.e. for the lasso the objective function is,

$$\left\|\mathbf{d} - \mathbf{L}\frac{\hat{\mathbf{W}}^T}{||\hat{\mathbf{W}}||_2}\mathbf{d}\right\|^2 + \lambda||\hat{\mathbf{s}}||_1.$$

Of course we are now optimising for the weights rather than the parameter vector $\hat{\mathbf{s}}$, therefore since $\mathbf{W}$ is a matrix it is not clear how we would solve this for the lasso and other sparse methods. Furthermore, it is likely that this new objective function would not result in sparse solutions.

In addition to the deep sources, the sparse methods also have some issues with producing smooth solutions. The use of the ℓ_1 norm means that due to inconsistencies in the selection over time, we naturally recover estimates that are temporally spiky. This is especially true for the lasso, where due to its issues with correlated covariates, the chosen location of activity may jump around considerably. One simple way to address this would be to employ some form of averaging (for example a weighted moving average) to provide better temporal smoothness, however doing so also has a detrimental effect on MEG's strong temporal resolution (Uutela et al., 1999).

Within the MEG literature there have been a number of attempts to address this problem and to develop methods that provide the advantages of sparseness from the ℓ_1 norm whilst preserving some level of smoothness in the spatial and temporal aspects. Gramfort and Kowalski (2009), proposed a method using a mixed norm with multiple experimental conditions, which may relate to different cortical regions. The ℓ_1 norm is applied to the conditions while ℓ_2 norms are employed over space and time to give smooth solutions. This has some relation to the group lasso, where we ensure that groups of covariates are selected together (Yuan and Lin, 2006). Another approach utilising ℓ_1 minimisation is VESTAL/Fast-VESTAL (Huang et al., 2006, 2014). The VESTAL approach begins by applying singular value decomposition to the leadfield matrix. i.e. $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, then VESTAL first looks to solve

$$\min(\mathbf{w}^T||\mathbf{s}||_1) \text{ subject to } \mathbf{\Sigma}_{nl}\mathbf{V}_{nl}^T\mathbf{s} \cong \mathbf{U}_{nl}^T\mathbf{d},$$

where $\mathbf{w}$ is an $n \times 1$ weight vector and the subscript nl refers to the nl largest singular values and corresponding singular vectors. In order to remove the ℓ_1 norm whilst satisfying the non-negative requirements, the $p \times p$ diagonal sign matrix $\mathbf{\Omega}$ is introduced containing the signs for each dipole component,

$$\min(\mathbf{w}^T\mathbf{s}_+) \text{ s.t } \mathbf{\Sigma}_{nl}\mathbf{V}_{nl}^T\mathbf{\Omega}\mathbf{s}_+ \cong \mathbf{U}_{nl}^T\mathbf{d},$$

where $\mathbf{s}_+$ is the non-negative strength vector. VESTAL then uses the fact that the recorded magnetic fields are linear functions of distributed sources in order to provide smooth time courses. Taking the singular value decomposition of the data matrix, $\mathbf{d} = \mathbf{U}_d\mathbf{\Sigma}_d\mathbf{V}_d^T$, the estimated time courses are then projected towards the right singular vectors to ensure that the recorded data and reconstructed source time courses share the same temporal information. The VESTAL solution is then,

$$\hat{\mathbf{s}}_{VESTAL} = \hat{\mathbf{s}}\mathbf{P},$$

where the projection matrix $\mathbf{P} = \mathbf{V}_d\mathbf{V}_d^T$ (Huang et al., 2006). The Fast-VESTAL method (Huang et al., 2014) looks to speed up the procedure by removing the time dependency at the beginning of the method.

These methods will not be implemented in this thesis, however they have been shown to perform well in MEG simulations and real data. The VESTAL method produced much smoother estimates than the conventional ℓ_1 norm approach with none of the spikiness of the lasso time courses (Huang et al., 2006). Furthermore, the Fast-VESTAL modification was able to correctly reconstruct the correlation structure between multiple correlated sources and provide much more spatially exact source estimation than the beamformer (Huang et al., 2014). The mixed norm approach of Gramfort and Kowalski (2009) overcame a number of the limitations of the standard ℓ_1 or ℓ_2 approaches as the method was less affected by the signal-to-noise ratio than the lasso, and lacked the spatial blurring of the minimum norm. In terms of Bayesian methods, the multiple sparse priors approach (Friston et al., 2008) was able to account for a much larger proportion of the variance of the data than either the unconstrained minimum norm or a more realistic alternative. Furthermore, it was superior in both spatial and temporal reconstruction.

In the chapters that follow, we will focus on measurement-error in MEG and methods for correcting sparse estimates from such error. Many of these methods are designed for use with linear regression and can be modified to include penalisation. There are a number of Bayesian approaches to measurement-error correction, however these tend to be computationally intensive and we are required to specify the full likelihood including the distribution of the forward model (which in our case is the leadfield, $\mathbf{L}$)(Carroll et al., 2006). Therefore,

going forward we will focus on approaches that are suitable for the sparse regression methods outlined in this chapter.

3

MEASUREMENT-ERROR AND THE CONDITIONAL SCORE

Consistent estimation of parameters in both linear and non-linear models is hugely dependent on the accuracy of the predictors. Regressing on an error contaminated model design matrix can have dramatic effects on a model's performance with respect to bias. Depending on the model construction and distribution of the covariates among other factors, measurement-error can result in biasing estimates towards zero (also known as attenuation), hiding true features, changing the signs of parameter estimates or exhibiting erroneous relationships between covariates (Fuller, 1987; Carroll et al., 2006). For example, in nutritional epidemiology, measurement-error in dietary measures results in attenuation in the estimated relative risk and severely affects the statistical power of such studies, which may obscure genuine relationships between diet and disease (Kipnis et al., 2003). Therefore, it is important that the predictors used are accurate.

Unsurprisingly, in many applications of statistical modelling, we cannot always guarantee precision in the specification of model predictors. This can occur either through measurement inaccuracies or in problems when modelling the forward solution is non-trivial. In such cases, given the estimation con-

sequences of measurement-error, it is important that we give some thought to model correction. This can then inform which features of the parameter space are actually important to the model and which are merely a result of measurement-error. This obviously has important implications for MEG problems, where measurement-error can be introduced into the leadfield matrix from a number of sources, either modelling or experimental (Lopez et al., 2012; Akalin Acar and Makeig, 2013). In the MEG setting, measurement-error models will allow us to account for an erroneous leadfield matrix. Therefore we now look to introduce methods for tackling measurement-error in linear models that we can adapt for MEG (and similar large dimensional) problems.

There are a number of possible approaches to take in measurement-error modelling, with parametric, non parametric, frequentist and Bayesian models all being employed to some extent (examples of these can be found in Carroll et al., 2006). The method that we will focus on in this chapter is a score function method known as conditional score. It has the advantage of being fairly simple in both programming and implementation as well as being consistent in its estimation without making assumptions about the distribution of the model design $\mathbf{X}$.

3.1 MEASUREMENT-ERROR IN THE LINEAR MODEL

We begin by outlining the linear model with the presence of measurement-error and explain how naive estimation under these circumstances results in biased estimates. We then look to introduce the conditional score method, which attempts to account for the measurement-error within the estimation.

The additive measurement-error model for linear regression is as follows;

$$Y_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{X}_i + \epsilon_i$$
$$\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i, \quad \text{for } i = 1, ..., n. \tag{3.1}$$

For the $1 \times n$ observation vector $\mathbf{Y}$, the $1 \times p$ parameter vector $\boldsymbol{\beta}^T$ and the $p \times n$ model matrix $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_n)^T$, the model has mean $E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \boldsymbol{\beta}^T\mathbf{X}$, and variance $\text{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{I}_n$, i.e. $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$. The observed model matrix with measurement-error is $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_u)$. The measurement-error covariance matrix is assumed to be known (or at least able to be estimated), for example, say $\boldsymbol{\Sigma}_u = \sigma_u^2\mathbf{I}_p$ (note: $\boldsymbol{\Sigma}_u$ is not required to be diagonal). The unknown parameters are summarised by $\Theta = (\beta_0, \boldsymbol{\beta}, \sigma^2)^T$. Note, we usually assume that $\mathbf{X}$ is independent of both the measurement-error and the additive model noise. We also assume that the two sources of error have no dependence on each other.

3.2 NAIVE ESTIMATE

Therefore, our only knowledge of the true model matrix $\mathbf{X}$ is through the error contaminated version $\mathbf{W}$. If we ignore the error in the matrix $\mathbf{W}$ and perform standard regression methods using the contaminated model matrix we obtain, what is known as, the *naive estimate*. It can be shown that the naive estimate suffers from bias (Fuller, 1987). For example in the case of the simple linear model with $p = 1$, with measurement-error variance σ_u^2 and predictor variance σ_x^2, for $n \to \infty$ the naive estimate $\hat{\beta}_{naive}$ has expectation,

$$\mathbb{E}(\hat{\beta}_{naive}) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\beta.$$

The standard OLS estimator of $\mathbf{Y}$ regressed on $\mathbf{X}$ can be shown to be,

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}}$$

where S_{AB} denotes the sample covariance between A and B. Now for the naive estimate we regress $\mathbf{Y}$ on $\mathbf{W}$ rather than $\mathbf{X}$. Hence,

$$\begin{aligned}\hat{\beta}_{naive} &= \frac{S_{WY}}{S_{WW}} \\ &= \frac{S_{XY} + S_{UY}}{S_{XX} + 2S_{UX} + S_{UU}}.\end{aligned}$$

Now, as $n \to \infty$, $\mathbb{E}(S_{UY}) \to 0$, $\mathbb{E}(S_{UX}) \to 0$, $\mathbb{E}(S_{XX}) \to \sigma_x^2$ and $\mathbb{E}(S_{UU}) \to \sigma_u^2$. Therefore, for large n,

$$\mathbb{E}(\hat{\beta}_{naive}) = \mathbb{E}\left(\frac{S_{XY}}{S_{XX} + S_{UU}}\right) = \frac{1}{\sigma_x^2 + \sigma_u^2}\mathbb{E}\left(\frac{S_{XX}S_{XY}}{S_{XX}}\right) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\mathbb{E}(\hat{\beta}).$$

Hence, using the fact that the OLS estimator $\hat{\beta}$ is unbiased,

$$\mathbb{E}(\hat{\beta}_{naive}) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\beta.$$

We call the ratio $\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$ the *reliability ratio* which is denoted λ_{RR}. The reliability ratio naturally lies in [0,1] and its value depends on the variance of the measurement-error. When σ_u^2 is small in comparison to σ_x^2 the bias will naturally also be small, however larger values will have a more dramatic impact on the accuracy of the naive estimator. Nevertheless in general, the naive estimator is a biased estimator in the presence of measurement-error. By comparison, the conditional score method looks to construct a consistent estimator for β under measurement-error.

Before reviewing the conditional score method in section 3.4 we will perform some simulation studies in order to assess the impact of measurement-error on sparse regression methods.

3.3 EFFECT OF MEASUREMENT-ERROR ON SPARSE REGRESSION

3.3.1 *Monte Carlo simulations*

We now investigate the effect of measurement-error on the sparse methods outlined in chapter 2. In particular we are interested in the effect on covariate selection when we increase the measurement-error. In the first simulation, we perform a Monte Carlo study for 250 repetitions.

At each repetition, we begin by simulating 500 parameters. All but 5 of these are equal to zero, the remaining 5 are simulated from $\mathrm{N}(0, 2^2)$. The model matrix $\mathbf{X}$, for 100 observations is simulated from a standard normal and the noise, ϵ, from $\mathrm{N}(0, 0.1)$. The measurement-error is then simulated from $\mathrm{N}_p(\mathbf{0}_p, \sigma_u^2 \mathbf{I}_p)$, where σ_u^2 is the measurement-error variance. In this simulation we use $\sigma_u^2 = 0$, 0.1, 0.2, 0.4, and 0.6. These measurement-error levels correspond to reliability ratios of 1, 0.91, 0.83, 0.71 and 0.63 respectively. In each simulation, we calculate the naive estimate for ridge, lasso, elastic net, square root lasso and PED as defined in section 2.1 (Note; going forward the "naive elastic net" refers to the elastic net estimate using the measurement-error contaminated $\mathbf{W}$, this differs from the previous term usage in section 2.1.4). The cross validated *glmnet* function was used to calculate the ridge, lasso and elastic net solutions, the square root lasso was fitted with the asymptotic penalty level using the *slim* function and PED was fitted with the sparsity threshold and penalisation level chosen via BIC from a grid of values with $C \in \{0.5, 0.75, 1, 1.25, 1.5, 1.7, 2, 2.2\}$ and 50 λ's between 0.001 and 1.5. Table 4 gives the means for the absolute sum of the bias, the number of true positives and the number of false positives

over the 250 Monte Carlo repetitions. The standard deviations are given in parenthesis.

Method	σ_u^2	Bias	True +	False +
Min norm	0	18.49 (6.56)	-	-
	0.1	18.17 (6.48)	-	-
	0.2	17.90 (6.49)	-	-
	0.4	17.10 (6.40)	-	-
	0.6	16.03 (6.41)	-	-
E-Net	0	1.10 (0.34)	4.78 (0.48)	36.38 (13.33)
	0.1	4.61 (2.17)	4.30 (0.77)	37.26 (19.62)
	0.2	5.84 (2.78)	4.11 (0.79)	34.30 (19.98)
	0.4	7.26 (3.65)	3.86 (0.86)	31.63 (22.12)
	0.6	7.79 (3.62)	3.60 (0.92)	28.30 (21.33)
Lasso	0	0.72 (0.26)	4.78 (0.46)	22.88 (13.08)
	0.1	3.44 (1.70)	4.29 (0.76)	24.55 (17.41)
	0.2	4.71 (2.31)	4.07 (0.81)	23.16 (16.67)
	0.4	6.21 (3.08)	3.79 (0.89)	22.72 (19.64)
	0.6	6.65 (2.93)	3.50 (0.93)	18.26 (14.82)
SRL	0	4.44 (2.70)	3.32 (0.80)	0 (0.06)
	0.1	5.49 (2.54)	2.70 (0.69)	0 (0)
	0.2	6.03 (2.59)	2.31 (0.68)	0 (0.06)
	0.4	6.62 (2.64)	1.86 (0.63)	0.01 (0.09)
	0.6	6.96 (2.71)	1.57 (0.67)	0 (0.06)
PED	0	0.61 (0.63)	4.18 (0.75)	4.22 (7.51)
	0.1	6.29 (2.49)	3.87 (0.84)	22.36 (6.80)
	0.2	9.72 (3.70)	3.81 (0.84)	29.10 (5.73)
	0.4	12.91 (4.75)	3.73 (0.87)	35.84 (5.16)
	0.6	14.61 (5.30)	3.55 (0.84)	40.34 (5.38)

Table 4: Bias, true positives and false positives. Standard deviation in parenthesis. 250 Monte Carlo simulations.

Table 5 summarises the proportion of the signal correctly attributed to the true locations. This was calculated by $\frac{|\hat{\boldsymbol{\beta}}_{S_0}|_1}{|\boldsymbol{\beta}^0_{S_0}|_1}$ where $\hat{\boldsymbol{\beta}}$ is the estimated parameters and $\boldsymbol{\beta}^0$ is the vector of the true parameters. The subscript S_0 indicates the subset containing only the true parameters.

Method	σ_u^2	0	0.1	0.2	0.4	0.6
Min norm	Mean	0.08	0.07	0.07	0.06	0.05
	Sd	0.01	0.01	0.01	0.02	0.02
E-net	Mean	0.94	0.70	0.58	0.43	0.33
	Sd	0.07	0.07	0.07	0.07	0.07
Lasso	Mean	0.95	0.75	0.63	0.47	0.37
	Sd	0.06	0.07	0.069	0.07	0.07
SRL	Mean	0.45	0.30	0.23	0.15	0.10
	Sd	0.22	0.14	0.12	0.08	0.07
PED	Mean	0.95	0.68	0.54	0.40	0.31
	Sd	0.08	0.08	0.08	0.07	0.06

Table 5: Proportion of signal detected at true locations, $\frac{|\hat{\boldsymbol{\beta}}_{S_0}|_1}{|\boldsymbol{\beta}^0_{S_0}|_1}$.

From Table 4, the lasso appears to give the best performance as it has the lowest bias once measurement-error is introduced and has a good number of true positives. The elastic net gives very similar results to the lasso, but does have far more false positives. Given the nature of the simulation, this is fairly expected as the addition of the ridge penalty in the elastic net will result in more non-zero covariates. When the covariates are uncorrelated, the grouping property is redundant and we can get additional false positives included. The square root lasso is obviously too sparse at the level chosen as it has a relatively low average

number of true positives. On the other hand, the identification of unimportant covariates is very strong as there are no false positives. The performance of the square root lasso is poor even in the absence of measurement-error, therefore the asymptotic penalty value is not appropriate for the values of n and p in these simulations. In future simulations it will be preferable to choose from a series of penalty values using BIC or cross validation.

PED gives very good results when there is no measurement-error, performing well in terms of bias and both true and false positives. However, the performance is affected dramatically when measurement-error is introduced. The true positive rate remains on a similar level to the lasso, although the other measures of the performance suffer. Closer analysis shows that as the measurement-error increases we get more extraneous variables included, the magnitude of which also increase. Hence the estimates are much noisier, resulting in larger bias. The effect on the true covariates is investigated more in the next table as we look at the proportion of signal correctly attributed to the true parameters.

Table 5 shows that in the absence of measurement-error the lasso, elastic net and PED have similarly low bias for the true positives. Since the square root lasso with asymptotic penalty has chosen too sparse a model, its performance is much weaker. For the lower levels of measurement-error, the proportion of signal at the true covariates is still fairly good, with the lasso performing the best. Under larger measurement-error the estimates are much weaker. These results also support the earlier conclusions about the affect on PED estimates; that the degeneration of the estimate as measurement-error is introduced is more due to the inclusion of extraneous variables than the impact on the true covariates.

Unfortunately, some of the standard deviations in Tables 4-5 are very high. This is particularly the case for the bias and false positives. This means there is a large amount of uncertainty about these results. The amount of variability is mainly due to the number of Monte Carlo samples included. Performing 250 Monte Carlo repetitions for 5 methods at 5 different measurement-error levels took around a day and a half to calculate. The square root lasso and PED were by far the slowest methods and the bulk of the calculation time was spent on these methods. Although the calculation time for the naive estimates is high, it is not particularly extreme. However, we want to be able to compare the naive results with the estimates from measurement-error methods. These can be much slower to calculate, therefore increasing the number of samples significantly would be prohibitive. Further discussion of this is found at the end of section 7.1.

3.3.2 *MEG simulations*

We also want to look at the effect of measurement-error on our methods in the context of an MEG problem. The next simulation follows a similar form to the single source MEG simulations in section 2.3. We have a single, strength 5, source and a noise variance of 5. Here 500 time samples are simulated. The number of observations is equal to $n = 270$ and there are $p = 556$ parameters. We calculate the measurement-error free estimates and simulate measurement-error with variance σ_u^2=0.5, 1, 2, 5 and 10. The naive estimates are then calculated with the simulated error added to the original leadfield matrix.

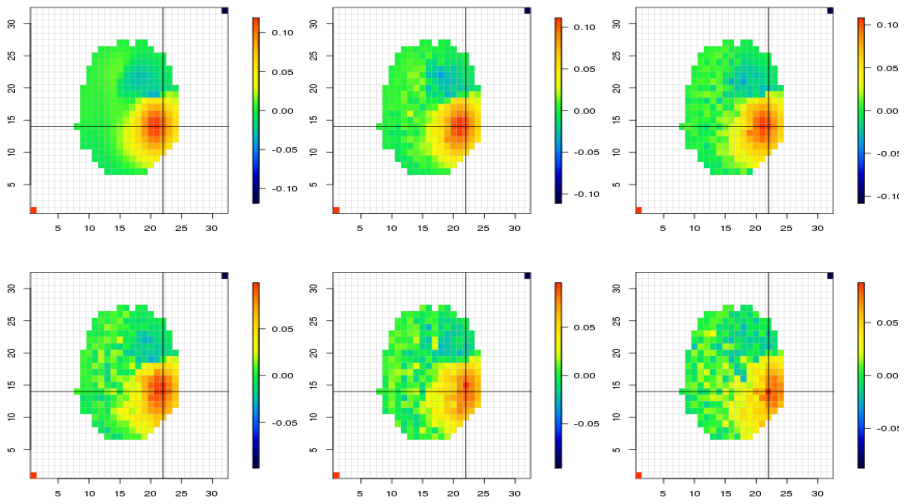

Figure 20: Effect of measurement-error on min-norm estimate, $\sigma_u^2 = 0$ (i.e. no measurement-error), $\sigma_u^2 = 0.5$, $\sigma_u^2 = 1$, $\sigma_u^2 = 2$, $\sigma_u^2 = 5$, $\sigma_u^2 = 10$. Images centred around zero (baseline = -max(pixel value)).

Table 6 gives the sum of the bias, standard deviation and root mean square error respectively for each method and each level of measurement-error. Figs. 20 - 24 show the change in the mean estimates as we increase the measurement-error.

Method	σ_u^2	0	0.5	1	2	5	10
Min norm	Bias	11.62	11.42	11.37	11.07	10.88	10.22
	Sd	0.73	1.26	1.36	1.52	1.51	1.44
	RMSE	11.85	11.92	11.87	11.67	11.43	10.76
E-Net	Bias	8.81	10.68	10.31	10.35	10.63	9.88
	Sd	2.37	7.31	6.54	6.13	5.17	4.34
	RMSE	10.04	14.93	14.07	13.75	13.29	12.07
Lasso	Bias	2.16	10.87	10.41	10.44	10.71	9.92
	Sd	3.26	8.35	7.26	6.74	5.53	4.58
	RMSE	4.26	15.78	14.64	14.24	13.61	12.25
SRL	Bias	9.21	9.20	9.27	8.78	9.92	8.09
	Sd	0.81	2.52	2.69	2.79	2.39	1.58
	RMSE	9.34	9.96	10.11	9.76	10.00	8.60
PED	Bias	9.71	14.88	13.60	13.32	12.43	14.03
	Sd	1.57	22.81	19.37	16.11	11.53	13.21
	RMSE	10.33	31.12	27.38	24.25	19.80	21.97

Table 6: Bias, standard deviation and RMSE for sparse methods under measurement-error.

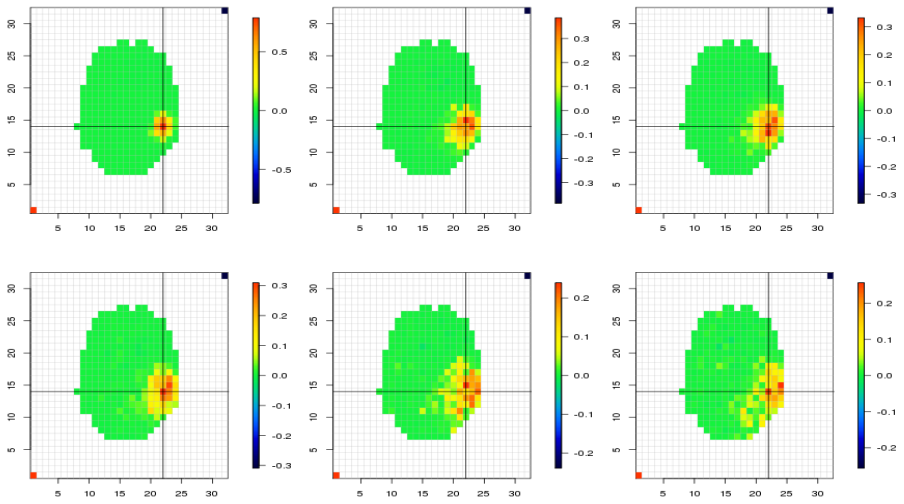

Figure 21: Effect of measurement-error on elastic net estimate, $\sigma_u^2 = 0$ (no measurement-error), $\sigma_u^2 = 0.5$, $\sigma_u^2 = 1$, $\sigma_u^2 = 2$, $\sigma_u^2 = 5$, $\sigma_u^2 = 10$.

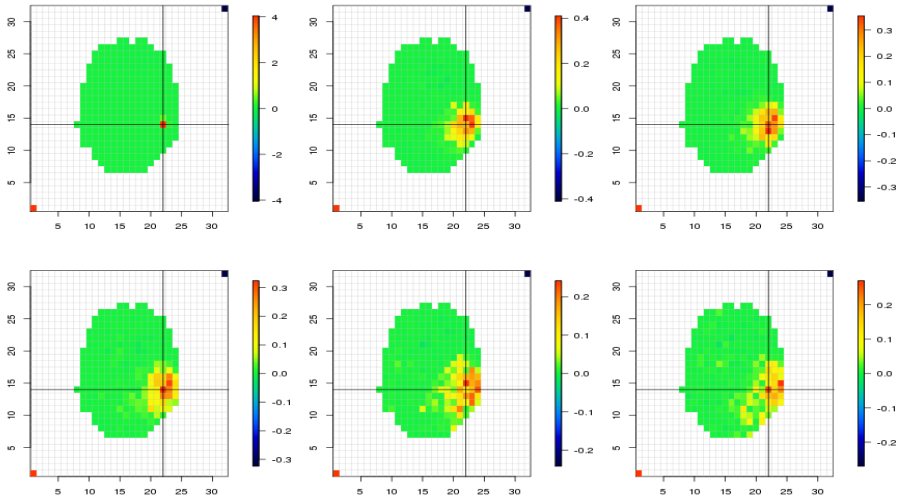

Figure 22: Effect of measurement-error on lasso estimate, $\sigma_u^2 = 0$ (no measurement-error), $\sigma_u^2 = 0.5$, $\sigma_u^2 = 1$, $\sigma_u^2 = 2$, $\sigma_u^2 = 5$, $\sigma_u^2 = 10$.

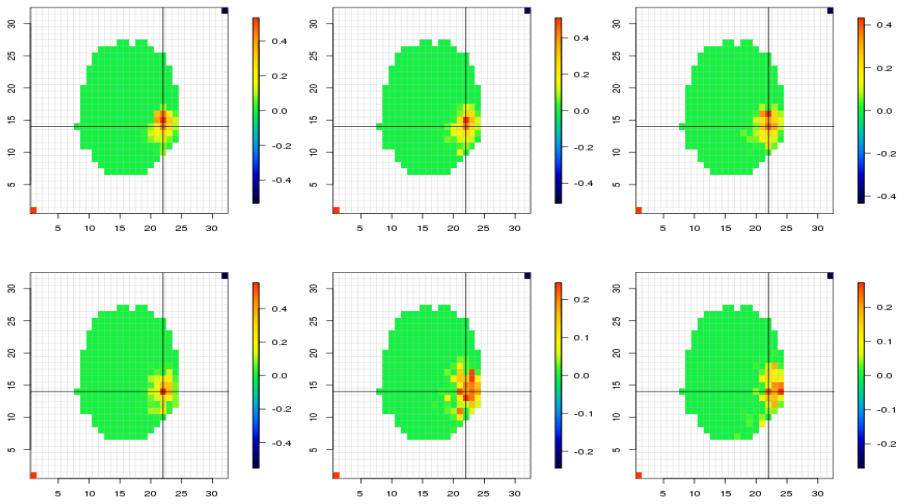

Figure 23: Effect of measurement-error on SRL estimate, $\sigma_u^2 = 0$ (no measurement-error), $\sigma_u^2 = 0.5$, $\sigma_u^2 = 1$, $\sigma_u^2 = 2$, $\sigma_u^2 = 5$, $\sigma_u^2 = 10$.

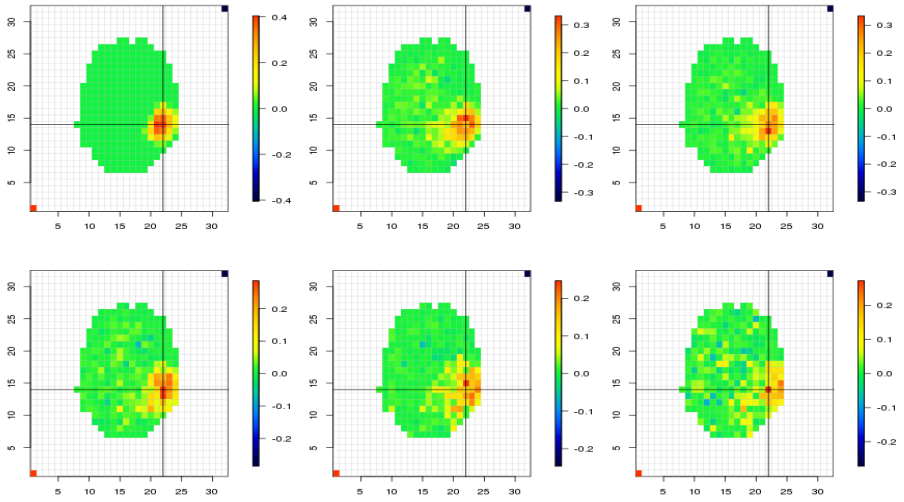

Figure 24: Effect of measurement-error on PED estimate, $\sigma_u^2 = 0$ (no measurement-error), $\sigma_u^2 = 0.5$, $\sigma_u^2 = 1$, $\sigma_u^2 = 2$, $\sigma_u^2 = 5$, $\sigma_u^2 = 10$.

The minimum norm estimate is fairly robust, even to the larger measurement-error levels. The main area of activity shows very little difference from the measurement-error free estimate. The non-source locations (green and light blue areas) do become noisier though. The elastic net sees some dispersal

of the source activity and for the higher measurement-error it places more activity on the edge of the slice. The same can be seen for the lasso, in fact when measurement-error is introduced, there is very little to separate the lasso and elastic net estimates. The square root lasso performs well and is largely unchanged for $\sigma_u^2 \leq 2$. The PED estimate is perhaps the most drastically affected and it becomes very noisy even for lower measurement-error. Whilst it still manages to identify the important areas of activity, the strength of the activity also drops considerably.

3.4 CONDITIONAL SCORE OVERVIEW

3.4.1 *The score function*

Before we describe the conditional score method, we must introduce the idea of the score function (see Pflug, 2002). A score function is defined as the first derivative of the log likelihood. Suppose we have a multiple linear regression model with mean $\beta_0 + \boldsymbol{\beta}\mathbf{X}$ and variance σ^2. To begin with, we assume there is no measurement-error and we write $(\beta_0, \boldsymbol{\beta})^T = \Theta$. The log likelihood is then,

$$\begin{aligned} l(\Theta, \sigma^2 | \mathbf{X}_i, Y_i) &= -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(Y_i - \left(1, \mathbf{X}_i^T\right)\Theta\right)\left(Y_i - \left(1, \mathbf{X}_i^T\right)\Theta\right)^T \\ &= -\log\sigma - \frac{1}{2\sigma^2}\left(Y_i^2 - \left(1, \mathbf{X}_i^T\right)\Theta\Theta^T\begin{pmatrix}1\\ \mathbf{X}_i\end{pmatrix} - 2Y_i\Theta^T\begin{pmatrix}1\\ \mathbf{X}_i\end{pmatrix}\right) \end{aligned}$$

differentiating with respect to Θ,

$$\frac{\partial l}{\partial \Theta} = -\frac{1}{2\sigma^2}\left(2\left(1, \mathbf{X}_i^T\right)\Theta\begin{pmatrix}1\\ \mathbf{X}_i\end{pmatrix} - 2Y_i\begin{pmatrix}1\\ \mathbf{X}_i\end{pmatrix}\right)$$
$$\propto \left(Y_i - \left(1, \mathbf{X}_i^T\right)\Theta\right)\begin{pmatrix}1\\ \mathbf{X}_i\end{pmatrix}.$$

Similarly for σ^2,

$$\frac{\partial l}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\left(Y_i - \left(1, \mathbf{X}_i^T\right)\Theta\right)^2$$

setting to zero and rearranging, we have,

$$0 = \sigma^2 - \left(Y_i - \left(1, \mathbf{X}_i^T\right)\Theta\right)^2.$$

The score function with no measurement-error is therefore,

$$\Psi_{ls}(Y_i, \mathbf{X}_i, \Theta) = \begin{bmatrix} \left\{Y_i - \left(1, \mathbf{X}_i^T\right)\Theta\right\}\begin{pmatrix}1\\ \mathbf{X}_i\end{pmatrix} \\ (\frac{n-p}{n})\sigma^2 - \left\{Y_i - \left(1, \mathbf{X}_i^T\right)\Theta\right\}^2 \end{bmatrix},$$

where $\frac{n-p}{n}$ is a degrees of freedom correction, where n is the number of observations and p is the number of parameters in Θ (Carroll et al., 2006). We then use the score function to estimate Θ by solving the estimating equations,

$$\sum_{i=1}^{n} \Psi_{ls}(Y_i, \mathbf{X}_i, \Theta) = \mathbf{0}.$$

3.4.2 *Conditional scores*

Under measurement-error, we obviously do not know the true value of the matrix $\mathbf{X}$. Treating the $\mathbf{X}_i$'s as additional parameters to maximise via functional maximum likelihood estimation is infeasible due to the shear number of

unknowns (Neyman and Scott, 1948). Furthermore, even when feasible, the estimators are not generally consistent (Stefanski and Carroll, 1987). Therefore, we base our inference on the conditional likelihood whereby we are able to derive a consistent estimator for the unknown, measurement-error free, predictor matrix in order to remove any dependence on $\mathbf{X}$. The conditional score method builds on the work of Anderson (1970) into the consistency of conditional maximum likelihood estimators. The idea behind the method is that we can introduce a parameter-dependent sufficient statistic for the unknown true regressors, so that the conditional distribution no longer depends on $\mathbf{X}$. From this we can derive unbiased estimators of the parameters. The solutions to the conditional score functions are then the conditional maximum likelihood estimators. This sufficient statistic we denote $\boldsymbol{\Delta}$ and is defined as,

$$\boldsymbol{\Delta}_i = \mathbf{W}_i + Y_i \boldsymbol{\Sigma}_u \boldsymbol{\beta} / \sigma^2.$$

This is derived by considering the distributions of $\mathbf{Y}$ and $\mathbf{W}$. We have,

$$\begin{aligned} f(Y_i|\mathbf{X}_i, \Theta) &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \boldsymbol{\beta}^T\mathbf{X}_i)^T(Y_i - \beta_0 - \boldsymbol{\beta}^T\mathbf{X}_i)\right\} \\ f(\mathbf{W}_i|\mathbf{X}_i) &= \left((2\pi)^p|\boldsymbol{\Sigma}_u|\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{W}_i - \mathbf{X}_i)^T\boldsymbol{\Sigma}_u^{-1}(\mathbf{W}_i - \mathbf{X}_i)\right\}. \end{aligned}$$

The joint density function is therefore,

$$\begin{aligned} f(Y_i, \mathbf{W}_i|\mathbf{X}_i, \Theta) &= \left((2\pi)^{p+1}\sigma^2|\boldsymbol{\Sigma}_u|\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \boldsymbol{\beta}^T\mathbf{X}_i)^2\right. \\ &\quad \left. -\frac{1}{2}(\mathbf{W}_i - \mathbf{X}_i)^T\boldsymbol{\Sigma}_u^{-1}(\mathbf{W}_i - \mathbf{X}_i)\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0)^2 - \frac{1}{2}\mathbf{W}_i^T\boldsymbol{\Sigma}_u^{-1}\mathbf{W}_i\right\} \\ &\quad \times \exp\left\{\mathbf{X}_i^T\boldsymbol{\Sigma}_u^{-1}(\boldsymbol{\Sigma}_u\boldsymbol{\beta}Y_i/\sigma^2 + \mathbf{W}_i)\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}^T\mathbf{X}_i\mathbf{X}_i^T\boldsymbol{\beta}/\sigma^2 + \mathbf{X}_i^T\boldsymbol{\Sigma}_u^{-1}\mathbf{X}_i + 2\beta_0\boldsymbol{\beta}^T\mathbf{X}_i/\sigma^2\right)\right\} \\ &\propto g_1(Y_i, \mathbf{W}_i) \times g_2(Y_i, \mathbf{W}_i, \mathbf{X}_i) \times g_3(\mathbf{X}_i). \end{aligned}$$

The joint conditional density can therefore be expressed as the product of three functions, where g_1 depends only on the known data $\mathbf{Y}, \mathbf{W}$; g_3 is a function only of the unknown $\mathbf{X}$ and g_2 is a function that depends on $\mathbf{X}$ and also the data through $\boldsymbol{\Delta}_i = \mathbf{W}_i + Y_i\boldsymbol{\Sigma}_u\boldsymbol{\beta}/\sigma^2$. Therefore, by the Fisher-Neyman Factorisation theorem, treating Θ as known, $\boldsymbol{\Delta}_i$ is a sufficient statistic for $\mathbf{X}_i$ (Stefanski and Carroll, 1987; Du, 2012).

Given $\mathbf{X}_i$, the random variables Y_i and $\boldsymbol{\Delta}_i$ are jointly normally distributed conditional on $\mathbf{X}_i$. It then follows that the conditional distribution $f(Y_i|\mathbf{X}_i, \boldsymbol{\Delta}_i)$ is also normal and the conditional expectation and variance can be derived from the density. Given the Jacobian of the transformation to the joint density of $(Y_i, \boldsymbol{\Delta}_i)$ is equal to one and given

$$Y_i \sim N(\beta_0 + \boldsymbol{\beta}^T\mathbf{X}_i, \sigma^2)$$
$$\boldsymbol{\Delta}_i \sim N(\mathbf{X}_i + Y_i\boldsymbol{\Sigma}_u\boldsymbol{\beta}/\sigma^2, \boldsymbol{\Sigma}_u),$$

we can find the joint density function,

$$f(Y_i, \Delta_i|\mathbf{X}_i) = \left((2\pi)^{p+1}\sigma^2|\boldsymbol{\Sigma}_u|\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \boldsymbol{\beta}^T\mathbf{X}_i)^2\right\}$$
$$\times \exp\left\{-\frac{1}{2}\left(\boldsymbol{\Delta}_i - Y_i\boldsymbol{\Sigma}_u\boldsymbol{\beta}/\sigma^2 - \mathbf{X}_i\right)^T\boldsymbol{\Sigma}_u^{-1}\left(\boldsymbol{\Delta}_i - Y_i\boldsymbol{\Sigma}_u\boldsymbol{\beta}/\sigma^2 - \mathbf{X}_i\right)\right\}.$$

Now, we want the conditional distribution of Y_i dependent on $\boldsymbol{\Delta}_i$. We can calculate this using,

$$f(Y_i|\boldsymbol{\Delta}_i) = \frac{f(Y_i, \boldsymbol{\Delta}_i|\mathbf{X}_i)}{f(\boldsymbol{\Delta}_i|\mathbf{X}_i)}.$$

Therefore, we need the distribution $f(\boldsymbol{\Delta}_i|\mathbf{X}_i)$ which we can calculate by integrating $\mathbf{Y}_i$ out of the joint density,

$$f(\boldsymbol{\Delta}_i|\mathbf{X}_i) = \int_{-\infty}^{\infty} f(Y_i, \boldsymbol{\Delta}_i|\mathbf{X}_i)dY_i$$
$$\propto \left((2\pi)^p\sigma^2|\boldsymbol{\Sigma}_u|\right)^{-\frac{1}{2}}(1 + \boldsymbol{\beta}^T\boldsymbol{\Sigma}_u\boldsymbol{\beta}/\sigma^2)^{-\frac{1}{2}}\exp\{-\frac{1}{2\sigma^2}(\beta_0 + \boldsymbol{\beta}^T\mathbf{X}_i)^2\}$$
$$\times \exp\{-\frac{1}{2}(\boldsymbol{\Delta}_i\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Delta}_i + \mathbf{X}_i^T\boldsymbol{\Sigma}_u^{-1}\mathbf{X}_i + \mathbf{X}_i^T\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Delta}_i)\}.$$

Then the ratio of the density functions is,

$$\begin{aligned} f(Y_i|\boldsymbol{\Delta}_i) =& \frac{f(Y_i, \boldsymbol{\Delta}_i|\mathbf{X}_i)}{f(\boldsymbol{\Delta}_i|\mathbf{X}_i)} \\ =& \frac{\sqrt{1+\boldsymbol{\beta}^T\boldsymbol{\Sigma}_u\boldsymbol{\beta}/\sigma^2}}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left(Y_i^2(1+\boldsymbol{\beta}^T\boldsymbol{\Sigma}_u\boldsymbol{\beta}/\sigma^2)\right.\right. \\ & \left.\left. -2Y_i(\beta_0+\boldsymbol{\beta}^T\boldsymbol{\Delta}_i)\right)\right\} \\ \propto& \frac{\sqrt{1+\boldsymbol{\beta}^T\boldsymbol{\Sigma}_u\boldsymbol{\beta}/\sigma^2}}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1+\boldsymbol{\beta}^T\boldsymbol{\Sigma}_u\boldsymbol{\beta}/\sigma^2}{2\sigma^2}\left(Y_i-\frac{\beta_0+\boldsymbol{\beta}^T\boldsymbol{\Delta}_i}{1+\boldsymbol{\beta}^T\boldsymbol{\Sigma}_u\boldsymbol{\beta}/\sigma^2}\right)^2\right\}. \end{aligned}$$

Therefore we have the conditional expectation and variance,

$$\mathbb{E}(Y_i|\mathbf{X}_i, \boldsymbol{\Delta}_i) = \mathbb{E}(Y_i|\boldsymbol{\Delta}_i) = \frac{\beta_0+\boldsymbol{\beta}^T\boldsymbol{\Delta}_i}{1+\boldsymbol{\beta}^T\boldsymbol{\Sigma}_u\boldsymbol{\beta}/\sigma^2},$$

$$\mathrm{Var}(Y_i|\mathbf{X}_i, \boldsymbol{\Delta}_i) = \mathrm{Var}(Y_i|\boldsymbol{\Delta}_i) = \frac{\sigma^2}{1+\boldsymbol{\beta}^T\boldsymbol{\Sigma}_u\boldsymbol{\beta}/\sigma^2}.$$

From Stefanski and Carroll (1987), the *conditional score* under measurement-error is then,

$$\boldsymbol{\Psi}_{cond}(Y_i, \mathbf{W}_i, \Theta) = \begin{bmatrix} \{Y_i - \mathbb{E}(Y_i|\boldsymbol{\Delta}_i)\}\begin{pmatrix}1\\ \boldsymbol{\Delta}_i\end{pmatrix} \\ \sigma^2 - \frac{\{Y_i-\mathbb{E}(Y_i|\boldsymbol{\Delta}_i)\}^2}{\mathrm{Var}(Y_i|\boldsymbol{\Delta}_i)/\sigma^2} \end{bmatrix}$$

and it can be seen that $\mathbb{E}\{\boldsymbol{\Psi}_{cond}(Y_i, \mathbf{W}_i, \Theta)\} = \mathbf{0}$. Hence, from the conditional score, when $\boldsymbol{\Sigma}_u$ is assumed known, we can form the unbiased estimating equations to be solved for Θ,

$$\sum_{i=1}^{n} \boldsymbol{\Psi}_{cond}(Y_i, \mathbf{W}_i, \Theta) = \mathbf{0}.$$

3.5 ASYMPTOTIC PROPERTIES

The conditional score estimate for the parameters solves the estimating equation,

$$\sum_{i=1}^{n} \Psi_{cond}(Y_i, \mathbf{W}_i, \Theta) = \mathbf{0}.$$

As such, the conditional score estimate is an M-estimator and from the properties of M-estimators we can infer the large sample distribution. M-estimators can be defined as the minima obtained from a sum of functions of the data (Huber, 1964).

Provided that $\mathbb{E}\{\Psi_{cond}(Y_i, \mathbf{W}_i, \Theta)\} = \mathbf{0}$, under certain regularity conditions (see Huber, 1967), the estimator $\hat{\Theta}$ asymptotically follows a normal distribution with mean Θ and covariance,

$$n^{-1} A_n^{-1}(\Theta) B_n(\Theta) \{A_n^{-1}(\Theta)\}^T$$

where $A_n(\Theta)$ and $B_n(\Theta)$ are defined as,

$$A_n(\Theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left\{ \frac{\partial}{\partial \Theta^T} \Psi_{cond}(Y_i, \mathbf{W}_i, \Theta) \right\}$$
$$B_n(\Theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left\{ \Psi_{cond}(Y_i, \mathbf{W}_i, \Theta) \Psi_{cond}^T(Y_i, \mathbf{W}_i, \Theta) \right\}.$$

3.6 DERIVATIVES

3.6.1 *Beta*

In order to solve the estimating equation for $\boldsymbol{\beta}$ we look to minimise the following;

$$\underset{\beta}{\operatorname{argmin}} \left\{ f(\boldsymbol{\beta}) = \left(\sum_{i=1}^{n} \Psi_{cond}(Y_i, \mathbf{W}_i, \Theta) \right)^2 \right\}.$$

For the optimisation process it is therefore helpful to have the derivatives of $f(\boldsymbol{\beta})$. For brevity let us denote $\Psi_i = \Psi_{cond}(Y_i, \mathbf{W}_i, \Theta)$, the derivative is then

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = 2\left(\sum \Psi_i\right)\left(\sum \frac{\partial \Psi_i}{\partial \boldsymbol{\beta}}\right).$$

It is easy to see that, for the case where the intercept $\beta_0 = 0$, the conditional score (with $\mathbb{E}(Y_i|\mathbf{X}_i, \boldsymbol{\Delta}_i)$ substituted in) is $\Psi_i = Y_i {\boldsymbol{\Delta}_i}^T - \frac{\boldsymbol{\beta}^T \boldsymbol{\Delta}_i {\boldsymbol{\Delta}_i}^T}{1+\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\sigma^2}$ and the sufficient statistic $\boldsymbol{\Delta}_i = \mathbf{W}_i + Y_i \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\sigma^2$. The derivative is therefore, using the quotient rule,

$$\frac{\partial \Psi_i}{\partial \boldsymbol{\beta}} = Y_i^2 \boldsymbol{\Sigma}_u/\sigma^2 - \left\{ \frac{\left(1+\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\sigma^2\right)\left(\frac{\partial(\boldsymbol{\beta}^T \boldsymbol{\Delta}_i {\boldsymbol{\Delta}_i}^T)}{\partial \boldsymbol{\beta}}\right) - \left(2\boldsymbol{\Sigma}_u \boldsymbol{\beta}/\sigma^2\right)\left(\boldsymbol{\beta}^T \boldsymbol{\Delta}_i {\boldsymbol{\Delta}_i}^T\right)}{\left(1+\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\sigma^2\right)^2} \right\}.$$

We therefore need the derivative of the product $\boldsymbol{\beta}^T \boldsymbol{\Delta}_i {\boldsymbol{\Delta}_i}^T$. Since

$$\begin{aligned}\boldsymbol{\beta}^T \boldsymbol{\Delta}_i {\boldsymbol{\Delta}_i}^T =& \boldsymbol{\beta}^T \mathbf{W}_i {\mathbf{W}_i}^T + \left((\boldsymbol{\beta}^T \mathbf{W}_i Y_i)\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta} Y_i {\mathbf{W}_i}^T\right)/\sigma^2 \\ &+ \left(\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta} Y_i^2\right)\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u / \left(\sigma^2\right)^2,\end{aligned}$$

the derivative is then,

$$\begin{aligned}\frac{\partial\left(\boldsymbol{\beta}^T \boldsymbol{\Delta}_i {\boldsymbol{\Delta}_i}^T\right)}{\partial \boldsymbol{\beta}} =& \mathbf{W}_i {\mathbf{W}_i}^T + \left(\boldsymbol{\beta}^T \mathbf{W}_i Y_i \times \mathbf{I}\boldsymbol{\Sigma}_u + \mathbf{W}_i Y_i \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u\right)/\sigma^2 \\ &+ 2\left(\boldsymbol{\Sigma}_u \boldsymbol{\beta} Y_i {\mathbf{W}_i}^T\right)/\sigma^2 + \frac{Y_i^2}{\left(\sigma^2\right)^2}\left(2\boldsymbol{\Sigma}_u \boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta} \times \mathbf{I}\boldsymbol{\Sigma}_u\right) \\ =& \mathbf{D}. \end{aligned} \tag{3.2}$$

Overall, we have

$$\begin{aligned}\frac{\partial f}{\partial \boldsymbol{\beta}} =& 2\left(\sum_{i=1}^{n}\left(Y_i {\boldsymbol{\Delta}_i}^T - \frac{\boldsymbol{\beta}^T \boldsymbol{\Delta}_i {\boldsymbol{\Delta}_i}^T}{1+\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\sigma^2}\right)\right) \\ &\left(\sum_{i=1}^{n}\left\{ Y_i^2 \boldsymbol{\Sigma}_u/\sigma^2 - \frac{\left(1+\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\sigma^2\right)\mathbf{D} - \left(2\boldsymbol{\Sigma}_u \boldsymbol{\beta}/\sigma^2\right)\left(\boldsymbol{\beta}^T \boldsymbol{\Delta}_i {\boldsymbol{\Delta}_i}^T\right)}{\left(1+\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\sigma^2\right)^2} \right\}\right). \end{aligned} \tag{3.3}$$

3.6.2 *Sigma squared*

We also need to use the conditional score in order to estimate the noise variance σ^2. If we let $\phi = \sigma^2$, then we are looking to minimise,

$$\underset{\phi}{\operatorname{argmin}} \left\{ f(\phi) = \left(\sum_{i=1}^{n} \left(\phi - \frac{\{Y_i - \mathbb{E}(Y_i|\boldsymbol{\Delta}_i)\}^2}{\operatorname{Var}(Y_i|\boldsymbol{\Delta}_i)/\phi} \right) \right)^2 \right\}.$$

Substituting in $\mathbb{E}(Y_i|\boldsymbol{\Delta}_i)$ and $\operatorname{Var}(Y_i|\boldsymbol{\Delta}_i)$,

$$f(\phi) = \left(n\phi - \sum_{i=1}^{n} \left(Y_i - \frac{\boldsymbol{\beta}^T \left(\mathbf{W}_i + \boldsymbol{\Sigma}_u \boldsymbol{\beta} Y_i/\phi \right)}{1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\phi} \right)^2 \left(1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\phi \right) \right)^2,$$

the derivative is then,

$$\frac{\partial f(\phi)}{\partial \phi} = 2 \left(n\phi - \sum_{i=1}^{n} \left(Y_i - \frac{\boldsymbol{\beta}^T \left(\mathbf{W}_i + \boldsymbol{\Sigma}_u \boldsymbol{\beta} Y_i/\phi \right)}{1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\phi} \right)^2 \left(1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\phi \right) \right) \left(\frac{\partial \Psi_\phi}{\partial \phi} \right)$$

where,

$$\begin{aligned} \frac{\partial \Psi_\phi}{\partial \phi} = & \sum_{i=1}^{n} \left(1 - \left(Y_i - \frac{\boldsymbol{\beta}^T \left(\mathbf{W}_i + \boldsymbol{\Sigma}_u \boldsymbol{\beta} Y_i/\phi \right)}{1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\phi} \right)^2 \left(-\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\phi^2 \right) \right. \\ & - 2 \left(Y_i - \frac{\boldsymbol{\beta}^T \left(\mathbf{W}_i + \boldsymbol{\Sigma}_u \boldsymbol{\beta} Y_i/\phi \right)}{1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\phi} \right) \left(1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\phi \right) \\ & \left. \times \frac{\left(1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\phi \right) \boldsymbol{\beta}^T \left(\boldsymbol{\Sigma}_u \boldsymbol{\beta} Y_i/\phi^2 \right) - \boldsymbol{\beta}^T \left(\mathbf{W}_i + \boldsymbol{\Sigma}_u \boldsymbol{\beta} Y_i/\phi \right) \left(\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\phi^2 \right)}{\left(1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}/\phi \right)^2} \right) \end{aligned}$$

is the derivative of the second conditional score equation with respect to the noise variance $\phi = \sigma^2$.

3.7 ALTERNATIVE FORMULATION

For the conditional score method we require foreknowledge of the variance of the measurement-error $\boldsymbol{\Sigma}_u$. In many cases, it may be more realistic to assume that the ratio of the error variances $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_u/\sigma^2$ is known (or at least viable

to estimate beforehand). This formulation was considered by Stefanski and Carroll (1987). Under this assumption the estimating equation becomes,

$$f(\phi) = \left(n\phi - \sum_{i=1}^{n} \left(Y_i - \frac{\boldsymbol{\beta}^T \left(\mathbf{W}_i + \boldsymbol{\Sigma}\boldsymbol{\beta} Y_i \right)}{1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}} \right)^2 \left(1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} \right) \right)^2 . \tag{3.4}$$

It is often useful to divide the conditional score through by σ^2 in order to give a steeper curve close to zero and thereby ensure the optimisation does not underestimate the variance. This can be seen in Fig. 25. When this is done the estimating equation becomes

$$f(\phi) = \left(n - \sum_{i=1}^{n} \left(Y_i - \frac{\boldsymbol{\beta}^T \left(\mathbf{W}_i + \boldsymbol{\Sigma}\boldsymbol{\beta} Y_i \right)}{1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}} \right)^2 \left(1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} \right) / \phi \right)^2 , \tag{3.5}$$

and the derivative is now

$$\begin{aligned} \frac{\partial f(\phi)}{\partial \phi} =& 2 \left(-n + \sum_{i=1}^{n} \left(Y_i - \frac{\boldsymbol{\beta}^T \left(\mathbf{W}_i + \boldsymbol{\Sigma}\boldsymbol{\beta} Y_i \right)}{1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}} \right)^2 \left(1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} \right) / \phi \right) \\ & \times \left(\sum_{i=1}^{n} \left(Y_i - \frac{\boldsymbol{\beta}^T \left(\mathbf{W}_i + \boldsymbol{\Sigma}\boldsymbol{\beta} Y_i \right)}{1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}} \right)^2 \left(1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} \right) / \phi^2 \right) . \end{aligned}$$

We know from the conditional score that for the optimal ϕ, $f(\phi)$ from 3.4 is equal to zero. Furthermore, we know that $\phi > 0$. Therefore, we can divide the conditional score estimating equation for ϕ through and obtain the same optimal ϕ in 3.5. An example of this is shown in Fig.25.

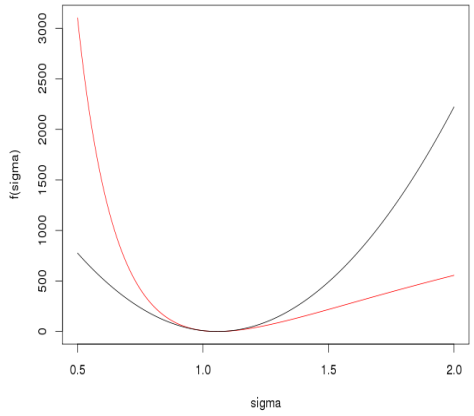

Figure 25: Conditional score estimating function plots for noise variance σ^2, with true value of $\sigma^2 = 1$. Estimating function from equation 3.4 (black) and equation 3.5 (red). Both curves have the same minimising value of σ.

3.8 EXAMPLE-SIMULATION

To demonstrate the conditional score method we perform a simple simulation. We simulate $p = 25$ parameters between -1 and 1 (15 of which were equal to zero) and simulate $n = 100$ observations. The model matrix is simulated from a standard normal distribution, $\mathbf{X} \sim \mathrm{N}_p(\mathbf{0}_p, \mathbf{I}_p)$ and the additive noise also followed a normal distribution with noise variance $\sigma^2 = 0.5$. For 500 Monte Carlo repetitions (here we can increase the size of our simulations over section 3.3 as the number of parameters is much smaller and we only need to calculate the naive and conditional score estimates for the ordinary least squares), measurement-error was added to the model matrix, $\mathbf{W} = \mathbf{X} + \mathbf{U}$ with $\mathbf{U} \sim \mathrm{N}_p(\mathbf{0}_p, \mathbf{\Sigma}_u = 0.3\mathbf{I}_p)$ and the naive estimate was calculated by regressing the data on the measurement-error contaminated design, $\mathbf{W}$. Assuming that the ratio $\mathbf{\Sigma} = \mathbf{\Sigma}_u/\sigma^2$ was known, but the noise variance was otherwise un-

known, the conditional score corrected estimate was obtained by successively optimising for the parameters and the estimated noise variance $\hat{\sigma}^2$ until convergence. The naive estimates were used as the starting point for this estimation. The reliability ratio equated to $\lambda_{RR} = 0.77$.

Figs. 26 and 27 show the means for both naive and conditional score estimates along with the true values of the parameters. The bias of the naive estimate can be seen, particularly in the larger parameters where we get a shrinkage type effect. The conditional score estimate is a clear improvement on the naive estimate and a number of the parameters have been almost perfectly corrected. The second plot however, shows that the variance of the conditional score estimates is larger than the naive case.

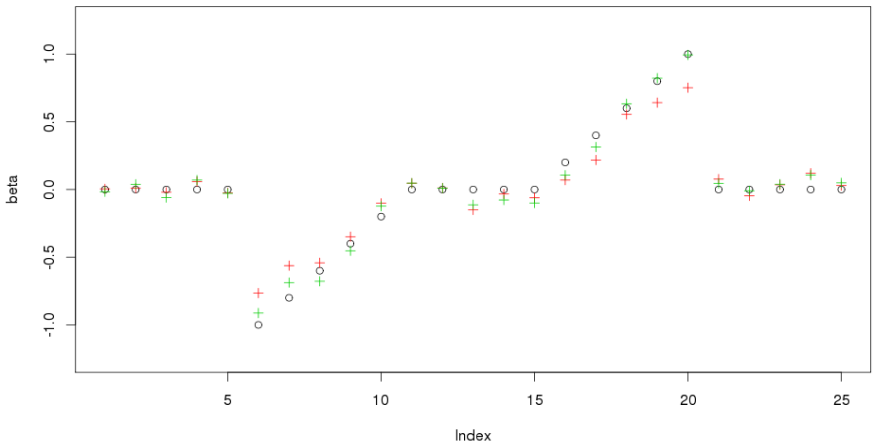

Figure 26: True parameter values (black) and mean estimates for naive (red) and conditional score (green) after Monte Carlo study.

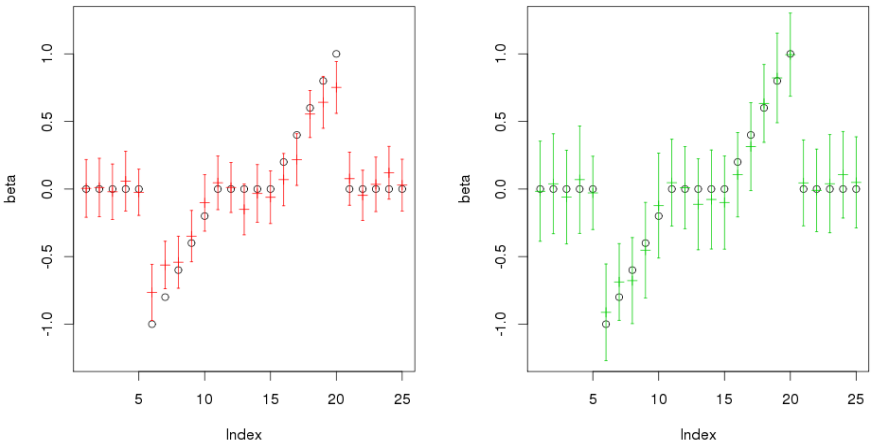

Figure 27: Mean estimates +/— 2 standard deviations for naive (red) and conditional score (green) after Monte Carlo study.

We can therefore see the trade off between bias and variance when estimating under measurement-error. This is further emphasised in Table 7, as despite the significant reduction in bias, the naive estimate does better in terms of the root mean square error. The variance can be reduced a little by increasing the number of observations n (usually to around $10 \times p$), but this will also improve the naive estimate (to a point). To some extent, higher variance in the measurement-error corrected estimate is unfortunately inescapable.

Estimate	sum(|Bias|)	sum(var)	sum(RMSE)
Naive	2.174	0.241	3.524
CS	1.458	0.702	4.485

Table 7: Summary of results for naive and conditional score estimates.

In the next chapter we will extend the conditional score method to cover penalised regression methods.

4

MEASUREMENT-ERROR: PENALISED CONDITIONAL SCORE

The conditional score method provides a way to correct the ordinary least squares naive estimate under measurement-error. However, in many cases the OLS method is no longer the most appropriate regression method to use. This becomes particularly important when we have a large number of parameters involved and we want to induce some sparsity or shrinkage in our estimates through the use of penalisation. In the previous chapter we have already noted that measurement-error affects sparse regression methods by introducing bias and confounding the selection. Therefore we want to look at applying the procedure of the conditional score to regression methods outside of the simple unconstrained linear models. In particular, we want to look to apply measurement-error methods to the ridge (minimum norm) and sparse regression methods. This will involve a modification of the conditional score method, which we will call penalised conditional score, to include the penalty term from the relevant penalised method. Before exploring these new methods, we first investigate the properties of the regression methods under measurement-error.

4.1 NAIVE RIDGE ESTIMATE

Ridge regression makes use of the ℓ_2 norm penalisation to shrink the coefficient estimates. This means that unlike the OLS estimate, the ridge estimate suffers from bias even in the case of no measurement-error. However, this is traded off by a reduction in variance. Recall the ridge estimate is;

$$\hat{\boldsymbol{\beta}}^{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2^2$$

Differentiating the likelihood for the ridge estimate and setting equal to zero we have,

$$\begin{aligned} \frac{\partial L^{ridge}}{\partial \boldsymbol{\beta}} &= 2(-\mathbf{X}^T)(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda \mathbf{I}\boldsymbol{\beta} \\ &\implies -\mathbf{X}^T\mathbf{Y} + (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})\boldsymbol{\beta} = 0 \end{aligned}$$

We then solve for $\boldsymbol{\beta}$ and arrange in terms of the ordinary least squares estimate $\hat{\boldsymbol{\beta}}^{ols}$,

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{ridge} &= (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}}^{ols} \\ &= (\mathbf{I} + \lambda \mathbf{I}(\mathbf{X}^T\mathbf{X})^{-1})^{-1}\hat{\boldsymbol{\beta}}^{ols} \\ &= \mathbf{Z}_x\hat{\boldsymbol{\beta}}^{ols}, \end{aligned}$$

where $\mathbf{Z}_x = (\mathbf{I} + \lambda \mathbf{I}(\mathbf{X}^T\mathbf{X})^{-1})^{-1}$ and assuming that $\mathbf{X}^T\mathbf{X}$ is invertible. If $\mathbf{X}$ is orthogonal then $\hat{\boldsymbol{\beta}}^{ridge} = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}^{ols}$. Now considering the ridge estimate under measurement-error, following the same reasoning as above,

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{ridge}_{naive} &= (\mathbf{W}^T\mathbf{W} + \lambda\mathbf{I})^{-1}\mathbf{W}^T\mathbf{Y} \\ &= (\mathbf{W}^T\mathbf{W} + \lambda\mathbf{I})^{-1}(\mathbf{W}^T\mathbf{W})(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Y} \\ &= (\mathbf{I} + \lambda\mathbf{I}(\mathbf{W}^T\mathbf{W})^{-1})^{-1}\hat{\boldsymbol{\beta}}^{ols}_{naive} \\ &= \mathbf{Z}_w\hat{\boldsymbol{\beta}}^{ols}_{naive},\end{aligned}$$

which implies that,

$$\mathbb{E}(\hat{\boldsymbol{\beta}}^{ridge}_{naive}) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\mathbf{Z}_w\mathbb{E}(\hat{\boldsymbol{\beta}}^{ols})$$

using the result for $\hat{\boldsymbol{\beta}}^{ols}_{naive}$ from section 3.2 and letting

$$\mathbf{Z}_w = (\mathbf{I} + \lambda\mathbf{I}(\mathbf{W}^T\mathbf{W})^{-1})^{-1}.$$

In order to derive the variance of the naive estimates, we write the estimators in terms of the unbiased least squares estimate. First recall that $\text{Var}(\hat{\boldsymbol{\beta}}^{ols}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ (Greene, 2008). The variance of the naive least squares estimate can be derived,

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{ols}_{naive} &= \lambda_{RR}\hat{\boldsymbol{\beta}}^{ols} \\ \implies \text{Var}(\hat{\boldsymbol{\beta}}^{ols}_{naive}) &= \lambda^2_{RR}\text{Var}(\hat{\boldsymbol{\beta}}^{ols}) \\ &= \sigma^2\lambda^2_{RR}(\mathbf{X}^T\mathbf{X})^{-1}.\end{aligned}$$

Therefore for the ridge estimate,

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{ridge} &= \mathbf{Z}_x\hat{\boldsymbol{\beta}}^{ols} \\ \implies \text{Var}(\hat{\boldsymbol{\beta}}^{ridge}) &= \sigma^2\mathbf{Z}_x(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Z}_x^T \\ \implies \text{Var}(\hat{\boldsymbol{\beta}}^{ridge}_{naive}) &= \text{Var}(\lambda_{RR}\mathbf{Z}_w\hat{\boldsymbol{\beta}}^{ols}) \\ &= \sigma^2\lambda^2_{RR}\mathbf{Z}_w(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Z}_w^T\end{aligned}$$

where λ_{RR} is the reliability ratio as defined in section 3.2.

4.2 PENALISED CONDITIONAL SCORE

We now look to generalise the conditional score method from section 3.4.2 for when we have penalisation on the $\boldsymbol{\beta}$ estimates. We might want to do this in order to constrain the parameter estimates by promoting sparsity or shrink them towards zero. Let $g(\boldsymbol{\beta}, \lambda)$ denote the penalisation— a function of beta that only depends on the regularisation constant λ. Since the score function is defined as the derivative of the log-likelihood, it follows that in our penalised conditional score we will require the derivative of the penalty rather than the penalty itself. In this context, the penalty can be seen as a prior function on $\boldsymbol{\beta}$ (for example consider that the ℓ_2 norm penalty can be represented with the Gaussian prior $\boldsymbol{\beta} \sim N(\mathbf{0}, \lambda^{-1}\mathbf{I})$). The penalised conditional score function will then be the standard conditional score with the addition of the first derivative of the penalty i.e,

$$\begin{aligned}\boldsymbol{\Psi}_{pen.cond}(Y_i, \mathbf{W}_i, \Theta) &= \boldsymbol{\Psi}_{cond}(Y_i, \mathbf{W}_i, \Theta) - \frac{\partial g(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}} \\ &= \{Y_i - E(Y_i|\boldsymbol{\Delta}_i)\} \begin{pmatrix} 1 \\ \\ \boldsymbol{\Delta}_i \end{pmatrix} - \frac{\partial g(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}}.\end{aligned}$$

The penalised optimising equation is then,

$$\underset{\beta}{\text{argmin}} \left\{ f_{pen}(\beta) = \left(\sum_{i=1}^{n} \boldsymbol{\Psi}_{cond}(Y_i, \mathbf{W}_i, \Theta) - \frac{\partial g(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}} \right)^2 \right\}.$$

Therefore, the derivative of $f_{pen}(\boldsymbol{\beta})$ is,

$$\begin{aligned}\frac{\partial f_{pen}}{\partial \boldsymbol{\beta}} =& 2 \left(\sum_{i=1}^{n} \left(Y_i \boldsymbol{\Delta}_i{}^T - \frac{\boldsymbol{\beta}^T \boldsymbol{\Delta}_i \boldsymbol{\Delta}_i{}^T}{1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta} / \sigma^2} \right) - \frac{\partial g(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}} \right) \\ & \times \left(\sum_{i=1}^{n} \left\{ Y_i^2 \boldsymbol{\Sigma}_u / \sigma^2 - \frac{(1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta} / \sigma^2)\mathbf{D} - (2\boldsymbol{\Sigma}_u \boldsymbol{\beta} / \sigma^2)(\boldsymbol{\beta}^T \boldsymbol{\Delta}_i \boldsymbol{\Delta}_i{}^T)}{(1 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta} / \sigma^2)^2} \right\} \right. \\ & \left. - \frac{\partial^2 g(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta} \boldsymbol{\beta}^T} \right),\end{aligned}$$

where $\mathbf{D}$ is defined in equation 3.2.

4.2.1 *Ridge regression*

For the case of the ridge penalty $g(\boldsymbol{\beta}, \lambda) = \lambda||\boldsymbol{\beta}||_2^2 = \lambda \sum_{j=1}^{p} \beta_j^2$. The conditional score penalisation is therefore

$$\frac{\partial g(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}} = 2\lambda\boldsymbol{\beta}$$

and the derivative is

$$\frac{\partial^2 g(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T} = 2\lambda\mathbf{1}^T.$$

4.2.2 *PED*

For the penalised Euclidean distance, we have a geometric mean of the ℓ_1 and ℓ_2 norms, $g(\boldsymbol{\beta}, \lambda) = \lambda\sqrt{||\boldsymbol{\beta}||_1||\boldsymbol{\beta}||_2}$.

The derivatives of the penalty become problematic when $\beta_j = 0$ since the derivative of the ℓ_1 norm penalty is undefined at zero. However, since the PED regression algorithm artificially sets parameters to zero after the model fitting, this should not be a problem. Therefore, for $\beta_j \neq 0$,

$$\frac{\partial g(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}} = \frac{\lambda}{2} \frac{\frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2}||\boldsymbol{\beta}||_1}{\sqrt{||\boldsymbol{\beta}||_1||\boldsymbol{\beta}||_2}} + \frac{\lambda}{2} \frac{sgn(\boldsymbol{\beta})||\boldsymbol{\beta}||_2}{\sqrt{||\boldsymbol{\beta}||_1||\boldsymbol{\beta}||_2}}.$$

For the second derivative, we split the first derivative into two quotients. Let $U_1 = \frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2}||\boldsymbol{\beta}||_1$, $U_2 = sgn(\boldsymbol{\beta})||\boldsymbol{\beta}||_2$ and $V = \sqrt{||\boldsymbol{\beta}||_1||\boldsymbol{\beta}||_2}$. Then the second derivative is equal to

$$\frac{\partial^2 g(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T} = \frac{\lambda}{2}\left(\frac{VU_1' - U_1V'}{V^2} + \frac{VU_2' - U_2V'}{V^2}\right).$$

The derivatives V', U_1' and U_2' are;

$$V' = \frac{1}{2} \frac{\frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2}||\boldsymbol{\beta}||_1}{\sqrt{||\boldsymbol{\beta}||_1||\boldsymbol{\beta}||_2}} + \frac{1}{2} \frac{sgn(\boldsymbol{\beta})||\boldsymbol{\beta}||_2}{\sqrt{||\boldsymbol{\beta}||_1||\boldsymbol{\beta}||_2}},$$

$$U_1' = \frac{||\boldsymbol{\beta}||_2 \left(||\boldsymbol{\beta}||_1 + sgn(\boldsymbol{\beta})\boldsymbol{\beta}\right) - \boldsymbol{\beta}||\boldsymbol{\beta}||_1 \frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2}}{||\boldsymbol{\beta}||_2^2},$$

$$U_2' = 2\delta(\boldsymbol{\beta})||\boldsymbol{\beta}||_2 + sgn(\boldsymbol{\beta})\frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2},$$

where $\delta(\boldsymbol{\beta})$ is the Dirac delta function, taking the value 0 everywhere except at $\beta_j = 0$, where its value is infinite. Since we specified that $\beta \neq 0$ we get $U_2' = sgn(\boldsymbol{\beta})\frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2}$.

Since the PED objective function involves the minimisation of the Euclidean distance between the data $\mathbf{Y}$ and the fitted values $\mathbf{X}\boldsymbol{\beta}$ rather than the squared distance, the derivative of the first term in the PED conditional score will be divided by the normalising factor $||Y_i - \mathbb{E}(Y_i|\boldsymbol{\Delta}_i)||_2$ i.e.,

$$\boldsymbol{\Psi}_{cond}(Y_i, \mathbf{W}_i, \Theta) = \frac{\{Y_i - \mathbb{E}(Y_i|\boldsymbol{\Delta}_i)\}\boldsymbol{\Delta}_i^T}{||Y_i - \mathbb{E}(Y_i|\boldsymbol{\Delta}_i)||_2} - \frac{\lambda}{2}\frac{\frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2}||\boldsymbol{\beta}||_1}{\sqrt{||\boldsymbol{\beta}||_1||\boldsymbol{\beta}||_2}} - \frac{\lambda}{2}\frac{sgn(\boldsymbol{\beta})||\boldsymbol{\beta}||_2}{\sqrt{||\boldsymbol{\beta}||_1||\boldsymbol{\beta}||_2}}. \tag{4.1}$$

4.2.3 *Elastic net, lasso and square-root lasso*

The elastic net, lasso and square-root lasso all involve the ℓ_1 norm penalisation in the objective function. This feature gives these methods the important sparsity property. However because of the discontinuity in the derivative, the ℓ_1 norm is not differentiable at zero. This presents problems for our penalised conditional score method, especially as the very value at which the derivative is not defined happens to be to be the value we expect many of the β_j's to be.

In order to overcome this problem, we will have to restrict our solutions to non-zero estimates (where the functions are differentiable) and enforce sparsity through thresholding in a similar way to PED. This takes the assumption that

for covariates that are equal to zero, the estimates are below some small value. Therefore for the elastic net penalty, with $\beta_j \neq 0$,

$$\frac{\partial g(\boldsymbol{\beta}, \lambda, a)}{\partial \boldsymbol{\beta}} = \lambda \left\{ (1-a)\text{sign}(\boldsymbol{\beta}) + 2a\boldsymbol{\beta} \right\}$$

where $a \in [0, 1]$ is the mixing between the ℓ_1 and ℓ_2 terms. The lasso penalty is obtained by setting $a = 0$. Additionally,

$$\frac{\partial^2 g(\boldsymbol{\beta}, \lambda, a)}{\partial \boldsymbol{\beta} \boldsymbol{\beta}^T} = 2\lambda a \mathbf{1}^T$$

so for the lasso penalty the second derivative is 0 for $\beta_j \neq 0$. It is worth noting that for the square-root lasso, the first term in the conditional score will be of the same form as PED (i.e. in equation 4.1) since it has the minimisation of a Euclidean distance in common.

4.3 SIMULATIONS

4.3.1 *Monte Carlo simulations*

Returning to the Monte Carlo simulations from section 3.3, we apply penalised conditional score to the naive PED estimates. The simulations were comprised of 500 Monte Carlo repetitions, each with $n = 100$ observations, $p = 500$ parameters and 5 non-zero parameters. Using the naive estimates as the starting point and the previously chosen value of the regularisation we optimise the conditional scores for $\boldsymbol{\beta}$ and σ^2 recursively until convergence. Since the conditional score algorithm is unable to set covariates to zero automatically, we threshold the estimates after the optimisation using the standard PED thresholding detailed in section 2.1.6. i.e.,

$$\frac{|\hat{\beta}_j|}{||\hat{\boldsymbol{\beta}}||_2} < \frac{C}{\sqrt{n}}$$

are set to zero, where C is a predetermined thresholding constant.

As we can see from the table below, the bias and number of false positives is hugely dependent on the value of C. We set the thresholding constant at 0.75 (the value originally obtained by BIC for the naive estimate), as well as 0.85 and 1 to investigate the effect on the corrected estimate.

		σ_u^2			
C		0.1	0.2	0.4	0.6
0.75	Bias	7.28 (4.21)	10.98 (4.11)	14.47 (4.49)	16.56 (4.60)
	TP	3.74 (0.88)	3.60 (0.94)	3.44 (0.90)	3.23 (0.94)
	FP	13.00 (10.68)	19.61 (9.23)	26.60 (8.01)	30.56 (6.70)
	Prop true	0.76 (0.11)	0.60 (0.11)	0.44 (0.10)	0.35 (0.08)
0.85	Bias	5.27 (2.71)	8.84 (3.11)	12.30 (4.04)	14.37 (4.27)
	TP	3.69 (0.90)	3.51 (0.95)	3.37 (0.92)	3.17 (0.97)
	FP	7.34 (6.07)	13.06 (5.37)	19.16 (5.05)	22.88 (4.19)
	Prop true	0.75 (0.12)	0.60 (0.12)	0.44 (0.10)	0.34 (0.08)
1	Bias	3.51 (1.64)	6.67 (2.61)	10.06 (3.88)	12.06 (4.44)
	TP	3.61 (0.94)	3.43 (0.98)	3.24 (0.98)	2.98 (0.97)
	FP	2.96 (2.63)	7.17 (3.15)	12.35 (4.52)	15.55 (5.20)
	Prop true	0.75 (0.12)	0.59 (0.12)	0.43 (0.10)	0.33 (0.09)

Table 8: Results for PED conditional score estimates for different measurement-error levels and using different threshold levels. Bias, true positives, false positives and proportion of signal detected at true locations. Simulations involved $n = 100$ observations and $p = 500$ parameters, of which 5 are non-zero.

Note that the values of the estimates $\hat{\beta}_i$ are otherwise unchanged for values above the threshold.

From the results it is evident that the penalised conditional score estimate has primarily rescaled the estimates. Both the true positives and false positives have increased in magnitude in the corrected fit. This can be seen by the fact that the proportion at the true locations and the bias have both increased in the lower threshold level. Increasing the threshold removes a number of the extraneous variables and reduces the bias to give better performance than the naive estimate. For the smaller levels of measurement-error the correction at the true locations is fairly significant, however even under the larger thresholds the bias and proportional strength are only on a par with the naive lasso.

4.3.2 *MEG simulations*

We simulated a group of 5 sources on a single slice as before for 500 time points. A group of sources was chosen in order to assess whether the selection of close locations is affected by measurement-error and the conditional score correction. Random noise was added to the data with variance 3 and the data was centred over time. Measurement-error was then simulated with mean 0 and variance $\sigma_u^2 = 3$. This was then added to the leadfield matrix to give the measurement-error contaminated leadfield. The naive estimates were then calculated using the erroneous leadfield. For the minimum norm, cross validation was used to choose the penalisation level. We used two methods to choose the relevant penalty and threshold levels for PED; BIC and the theoretical driven ξ method. We then implemented the penalised conditional score methods using the naive estimates and penalisation levels as our starting point in each case. Figs. 28 and 29 give the means and standard deviations of the naive and corrected estimates respectively.

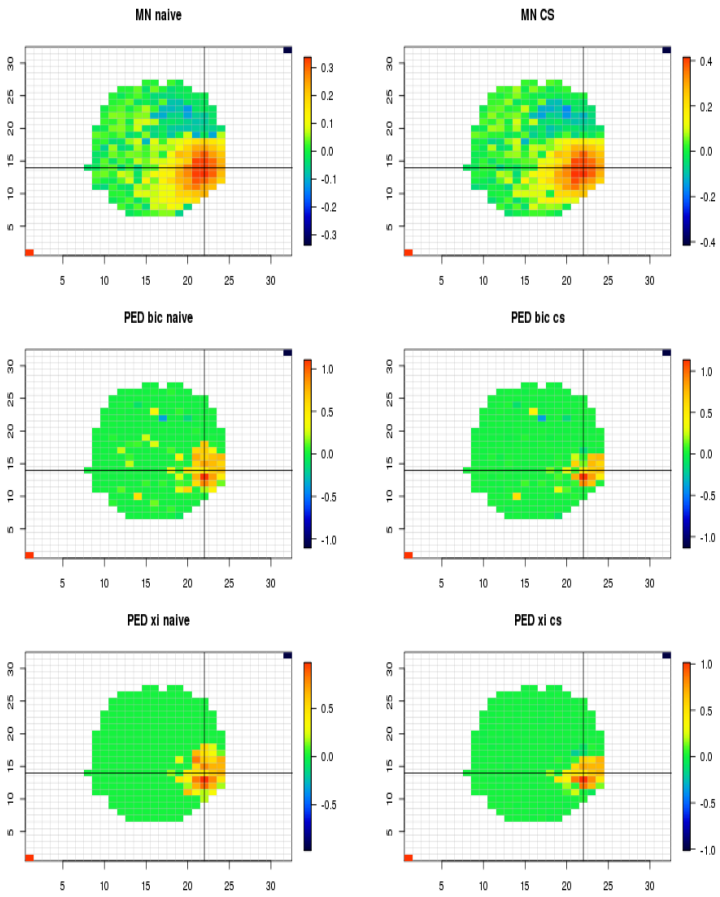

Figure 28: Mean of naive and conditional score corrected estimates for minimum-norm (ridge) and PED (bic and ξ).

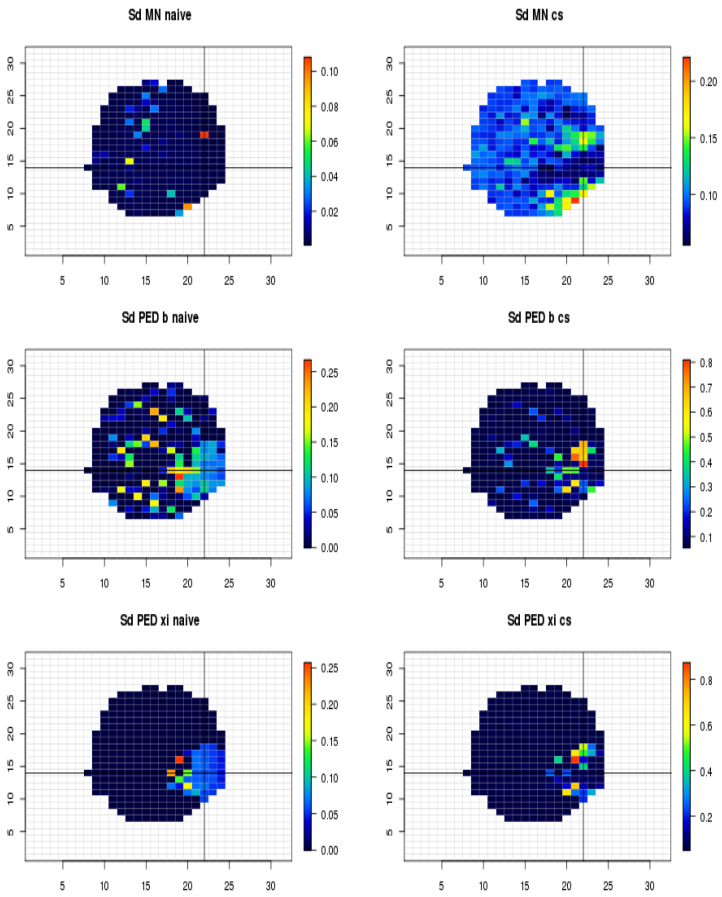

Figure 29: Standard deviation of naive and conditional score corrected estimates for minimum-norm (ridge) and PED (bic and ξ).

From Fig. 28 we see that for the ridge estimate the conditional score has primarily rescaled the naive estimate. The general pattern of the naive and conditional score estimates remain the same save for a few peripheral points of the main block of activity that have moved towards zero. We see a greater impact on the mean of the PED estimates where a number of false positives have seemingly been removed (or at least suppressed) and the activity in the conditional score estimates is generally less spread out. Unfortunately, in the case of the PED (BIC) conditional score estimate, a number of true positives have also been excluded. Therefore, the conditional score may have trouble selecting groups of closely located sources. Fig. 29 shows particularly high

variance on a number of the pixels that have been reduced by the conditional score estimate. This may indicate that there has been some sign change on one of the estimated components of the pixel strength.

4.4 POST SPARSE APPROACH

The conditional score as it stands is somewhat limited when it comes to sparse methods. To start with, the increased dimensionality that we almost inevitably see when discussing sparsity drastically complicates the optimisation required in the conditional score method. More specific to the sparse methods, however, the non-differentiability of the ℓ_1 norm is problematic. Since the derivative of the ℓ_1 norm is undefined at zero, the very value that the sparse methods set many of the parameters to, the penalised conditional score framework that we detailed above is no longer appropriate. For the PED we were able to work around this limitation as the sparsity is informed through a threshold after the model fitting, however the other sparse methods return truly sparse estimates.

An alternative method for adapting the conditional score for the sparse methods is to apply the conditional score only on the chosen non-zero parameters. In essence then, this *post sparse* application of the measurement-error correction uses the sparse fitting itself as a variable selection tool. Procedurally this means;

1. Fit the sparse method.

2. Perform ordinary least squares (OLS) or ridge regression on chosen non-zero predictors.

3. Using the chosen model from (2) as the starting "naive" estimate, apply conditional score method.

The use of a retrospective OLS fit over a sparse method is not without precedent. Belloni et al. (2011), employed an OLS fit over both the lasso and square root lasso. In both these cases the post fit performed well in comparison to the standard methods for a range of noise variances. Furthermore, the two-step method in the PED (Vasiliu et al., 2014), whereby we refit the method over the reduced parameter space, can be seen as a variation of this principle.

There are a number of advantages of this approach. Since the reduced problem is frequently no longer ill-posed, we have the advantage of easier computation in the conditional score. In addition, since OLS is now feasible, the subsequent estimate will not suffer from the bias present in penalised methods. The computational improvements translate into much quicker evaluating algorithms. However, from the measurement-error perspective, there are also a number of limitations. Primarily, since this approach only inputs the reduced parameter space into the conditional score step, parameters that are set to zero in the initial sparse fit will remain as zero throughout. When the number of true positives and true negatives are high there are no issues, however false negatives are particularly problematic. Therefore we need to be sure that the sparse methods are able to select the important parameters, even under significant measurement-error, as they will not be "corrected" into the model by the conditional score.

4.5 CONCLUSIONS

We have seen in chapter 3 that the inclusion of measurement-error in the model affects the performance of sparse regression methods. In the presence of measurement-error, sparse estimates suffer from attenuation bias and a deterioration in selection ability. In MEG type simulations, the estimates become increasingly noisy and more dispersed. The conditional score method represents one approach to the correction for the effects of measurement-error and in the $n > p$ case reduces the attenuation in the naive estimate considerably.

In this chapter we introduced penalisation into the conditional score method in order to extend the approach to cover larger dimensional problems and allow us to produce sparse conditional score estimates. The implementation of the PED conditional score in simulations shows that the amount of attenuation bias is reduced in the corrected estimate. However, investigation into the thresholding of the parameter estimates suggests that the false positives are also rescaled, which results in an increase of the overall bias. Additionally, for large dimensional problems the optimisation of the conditional score will become increasingly difficult. Therefore, we proposed the idea of reducing the dimensions by only using parameters that were selected by the naive estimate and fitting the ridge penalised conditional score on the remaining parameters. This process has some relation to thresholding functions like the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), as the smaller parameters are thresholded according to an ℓ_1 penalty and the larger parameters have a relaxed penalisation applied. The "post sparse" conditional score is applied to simulations in chapter 7.

5

SIMEX

Another method for measurement-error correction comes in the form of the SIMEX method. Cook and Stefanski (1994) (also see Stefanski and Cook, 1995), developed a simulation based method that looks to relate the covariate estimate with the level of measurement-error. It involves a two stage procedure, from which the method gets its name. The method begins with a *simulation* step involving model fitting, and is concluded with *extrapolation*. Hence we have the simulation-extrapolation or SIMEX algorithm.

5.1 METHOD

Suppose our estimation is based on the data $\mathbf{X}$ and we are only able to observe the measurement-error contaminated data matrix,

$$\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$$

for $i = 1, ..., n$ and $\mathbf{U}_i \sim N(\mathbf{0}_p, \sigma_u^2 \mathbf{I}_p)$. We assume that the measurement-error variance, σ_u^2, is known, or is at least able to be well estimated.

The simulation step of SIMEX relies on the basis that, whilst we are not able to reduce the variance of $\mathbf{W}$ through additional noise, we are able to study the

way the estimation is affected by increasing the measurement-error. With this in mind, the simulation step of SIMEX begins by simulating a number of new datasets that have increasing measurement-error,

$$\mathbf{W}_{i,b} = \mathbf{W}_i + \sqrt{\zeta}\mathbf{U}_{i,b}$$

where $b = 1, ..., B$. B is the number of simulation samples and $\zeta \geq 0$ controls the level of additional measurement-error. Each $\mathbf{U}_{i,b}$ is an iid simulation from the same distribution as the original measurement-error. The process of increasing the measurement-error may seem counter-intuitive when we are trying to remove the error from the model, however the motivation for this procedure can be seen when we consider the variance of the data matrices. It is easy to see that;

$$\text{Var}(\mathbf{W}_i|\mathbf{X}_i) = \sigma_u^2\mathbf{I}_p$$

and

$$\text{Var}(\mathbf{W}_{i,b}(\zeta)|\mathbf{X}_i) = (1+\zeta)\sigma_u^2\mathbf{I}_p.$$

We can therefore conclude that we will be able to recover the uncontaminated $\mathbf{X}_i$ when we have a value of $\zeta = -1$. Of course we are unable to produce simulations with a negative variance, however we can use simulations with increasing error to model the estimate as a function of ζ from which we can extrapolate to -1.

With the re-measurements of the data matrix $\mathbf{W}_{i,b}$ the naive estimates are then computed. One of the advantages of the SIMEX method is that we are free to choose the estimation method at this point. The naive estimates for sample b and error level ζ are denoted $\hat{\beta}_b(\zeta)$. For each level of ζ, the mean estimate across B is calculated,

$$\hat{\beta}(\zeta) = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b(\zeta).$$

The mean $\hat{\beta}(\zeta)$ is then the sample mean of the estimates over a number of experiments where we have the measurement-error level controlled by ζ. When this is calculated for increasing measurement-error we are able to plot ζ against $\hat{\beta}(\zeta)$. Note that the original naive estimate using the observed data $\mathbf{W}$ corresponds to $\zeta = 0$. Using this relationship between ζ and $\hat{\beta}(\zeta)$ we are then able to use a suitable extrapolation in order to estimate the value of $\hat{\beta}(-1)$.

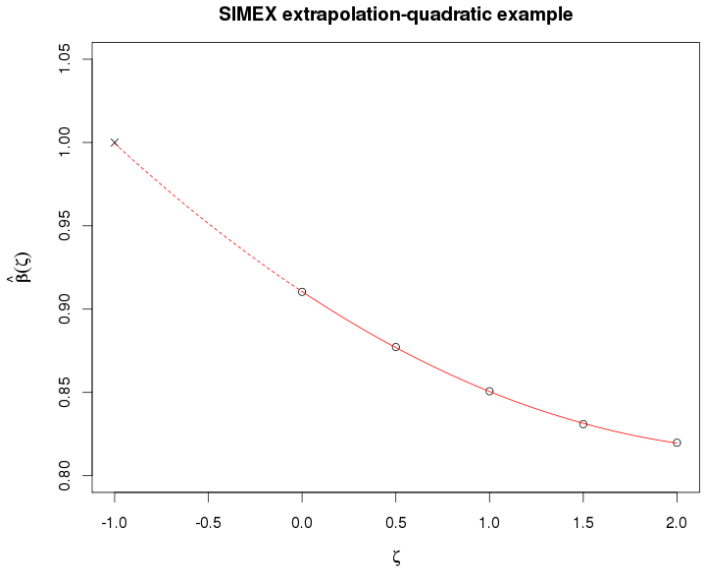

Figure 30: Example of SIMEX extrapolation using a quadratic extrapolant function. Increasing the value of ζ results in more biased estimates. The subsequent estimates for $\zeta \geq 0$ are used to produce an extrapolation to $\zeta = -1$ (dashed line).

Therefore, SIMEX uses the averages over the simulations to estimate the bias that occurs from increasing the measurement-error. The extrapolation is usually performed by using a suitable extrapolant function. A few common choices for the extrapolant function are,

Linear

$$\hat{\beta}(\zeta) = \gamma_0 + \gamma_1 \zeta$$

Quadratic

$$\hat{\beta}(\zeta) = \gamma_0 + \gamma_1\zeta + \gamma_2\zeta^2$$

Non-linear (rational linear)

$$\hat{\beta}(\zeta) = \gamma_0 + \frac{\gamma_1}{\gamma_2 + \zeta}.$$

The non-linear case can be solved by considering the three points $(\hat{\beta}(\zeta_0), \zeta_0)$, $(\hat{\beta}(\zeta_1), \zeta_1)$ and $(\hat{\beta}(\zeta_2), \zeta_2)$, where $\zeta_0 = 0, \zeta_1 = 1$ and $\zeta_2 = 2$ (Carroll et al., 2006). The covariates of the non-linear extrapolant are then given as,

$$\hat{\gamma}_0 = \frac{\hat{\beta}(\zeta_0)\left(\hat{\beta}(\zeta_2) - \hat{\beta}(\zeta_1)\right) - \hat{\beta}(\zeta_2)\left(\hat{\beta}(\zeta_1) - \hat{\beta}(\zeta_0)\right)}{\hat{\beta}(\zeta_2) - 2\hat{\beta}(\zeta_1) + \hat{\beta}(\zeta_0)}$$

$$\hat{\gamma}_1 = 2\frac{\left(\hat{\beta}(\zeta_1) - \hat{\beta}(\zeta_0)\right)\left(\hat{\beta}(\zeta_0) - \hat{\beta}(\zeta_2)\right)\left(\hat{\beta}(\zeta_2) - \hat{\beta}(\zeta_1)\right)}{\left(\hat{\beta}(\zeta_2) - 2\hat{\beta}(\zeta_1) + \hat{\beta}(\zeta_0)\right)^2}$$

$$\hat{\gamma}_2 = 2\frac{\hat{\beta}(\zeta_1) - \hat{\beta}(\zeta_2)}{\hat{\beta}(\zeta_2) - 2\hat{\beta}(\zeta_1) + \hat{\beta}(\zeta_0)}.$$

The extrapolation is then obtained by substituting these values and $\zeta = -1$ into the non-linear function. This extrapolant function is only really suitable under certain circumstances. For example, for parameters where the estimate does not change with increased measurement-error, it becomes obvious that the denominators of $\gamma_0, \gamma_1, \gamma_2$ approach zero and so they become undefined. Therefore for the non-linear function to be well defined we require some gradient in the line comparing ζ and $\hat{\beta}(\zeta)$. We can also see that there are problems when $0 < \hat{\gamma}_2 < 1$ as this coincides with a singularity in the region that we are extrapolating.

The first of these issues is of particular note for us, as the chances of parameters that are set to zero by our sparse methods being largely unaffected by measurement-error is fairly high. Of course if there is no gradient (as for a covariate that is set to zero at all measurement-error levels) then this will

imply that we have a situation where the measurement-error does not affect the particular covariate and $\hat{\beta}(-1) = \hat{\beta}(0)$.

SIMEX is an unusual method and at first it can appear to be counter-intuitive. However, the method has been shown to yield approximately unbiased and consistent estimates under the "correct" extrapolation and in the case of linear measurement-error modelling SIMEX is equivalent (or asymptotically equivalent) to the method-of-moments estimator (Cook and Stefanski, 1994). It is easy to see that the simulation stage of the method gives useful insight into the relationship between the measurement-error variance and the level of attenuation bias, however the subsequent extrapolation may be viewed with more wariness. Despite this, the method is not without some basis as the SIMEX method is actually related to the widely accepted jackknife method (Quenouille, 1956; Efron, 1982). Viewed under this framework, the extrapolation to $\zeta = -1$ is a modification of the reduced bias jackknife estimate (Stefanski and Cook, 1995).

5.1.1 *Simulations*

To test the SIMEX method we return to the simulation from chapter 3. For each of the sparse methods tested on the original dataset (lasso, elastic net, square-root lasso and PED) we use $\zeta \in \{0.5, 1, 1.5, 2\}$ and use $B = 100$ repetitions. The extrapolation step is then performed by fitting a quadratic model. The results from the 250 Monte Carlo simulations are summarised in Table 9. The naive equivalents are in Tables 4-5 in chapter 3. Comparing the tables, we can see that in all cases other than the square root lasso, the bias has increased. Additionally, the number of false positives has increased dramatically. On the

other hand, the number of true positives has remained largely unchanged or has improved. Another improvement has been in the proportion of signal attributed to the true positives.

Method	σ_u^2	0.1	0.2	0.4	0.6
Lasso	Bias	4.89 (3.34)	6.45 (4.38)	8.40 (6.25)	8.46 (5.24)
	TP	4.38 (0.74)	4.22 (0.77)	4.07 (0.84)	3.86 (0.88)
	FP	50.97 (24.43)	66.78 (26.64)	69.87 (27.81)	69.09 (29.80)
	Prop true	0.90 (0.10)	0.83 (0.10)	0.72 (0.13)	0.61 (0.13)
EN	Bias	6.76 (4.21)	8.35 (5.18)	10.47 (7.58)	10.95 (7.00)
	TP	4.40 (0.71)	4.31 (0.76)	4.10 (0.79)	3.89 (0.89)
	FP	87.41 (28.55)	92.91 (30.33)	94.88 (32.62)	94.23 (35.74)
	Prop true	0.86 (0.10)	0.78 (0.10)	0.66 (0.14)	0.55 (0.13)
SRL	Bias	4.81 (2.59)	5.21 (2.57)	5.81 (2.60)	6.25 (2.70)
	TP	2.94 (0.71)	2.64 (0.72)	2.32 (0.74)	1.98 (0.72)
	FP	0.11 (0.32)	0.20 (0.45)	0.33 (0.56)	0.33 (0.61)
	Prop true	0.40 (0.19)	0.34 (0.16)	0.26 (0.14)	0.20 (0.12)
PED	Bias	11.46 (4.10)	19.52 (6.96)	28.38 (10.27)	32.95 (12.04)
	TP	4.20 (0.79)	4.25 (0.80)	4.35 (0.70)	4.35 (0.74)
	FP	92.51 (16.03)	142.48 (22.76)	200.07 (32.11)	231.94 (37.62)
	Prop true	0.87 (0.13)	0.75 (0.13)	0.61 (0.13)	0.50 (0.12)

Table 9: Results for SIMEX estimates for different measurement-error levels. Bias, true positives, false positives and proportion of signal detected at true locations. Standard deviations in parenthesis.

Closer inspection reveals that the bias at the true locations has reduced in the SIMEX estimate, whereas the overall bias increase has come from the proliferation of small but non-zero parameter estimates. This is why the square root lasso with asymptotic penalty, which averages less than one false positive, has seen a reduction in the bias. The increase in false positives is an obvious

limitation in the SIMEX method coming from the extrapolation stage. For a covariate to be set to zero by the SIMEX estimate, it will require average values of zero at each level of ζ. The result is lots of small false positives that when summed result in a more substantial effect on the bias. Fig. 31 shows an example of this. Displayed are the true values, naive lasso estimates and SIMEX lasso estimates for one of the Monte Carlo simulations with measurement-error at $\sigma_u^2 = 0.1$. As demonstrated, the SIMEX correction for the important covariates is a major improvement over the naive equivalent. Unfortunately, there are a greater number of small extraneous variables and the false positives from the naive model have also become stronger.

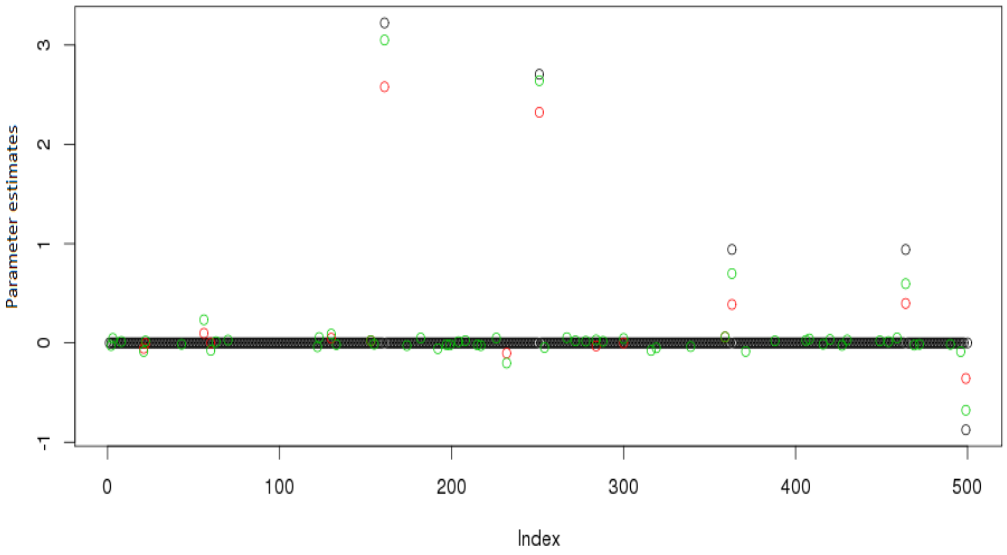

Figure 31: Naive (red) and SIMEX (green) estimates for the lasso when, $\sigma_u^2 = 0.1$. True values in black.

Therefore the main limitation of the SIMEX method for sparse models is that it relies on good selection in the naive estimates as it is unable to remove false positives or add false negatives back to the model. SIMEX is mostly able

to give some correction for attenuation, that is the shrinkage like effect that results from measurement-error. With this in mind we can either employ a thresholding to reduce the contribution of the extraneous variables, or we can apply SIMEX on the reduced problem by only using the covariates chosen by the naive estimate.

It is worth noting from the simulations that the naive estimates at the true positives are always closer to zero and there are no examples where the naive estimates returned are larger than the true values. The reason for this comes back to attenuation bias and the reliability ratio. If we recall, the reliability ratio for additive error is $\lambda_{RR} = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$ and since $\sigma_u^2 > 0$, the attenuation factor will be less than 1. Therefore, naive parameter estimates are shrunk towards zero.

5.2 MULTIPLICATIVE ERROR

We now look at how the SIMEX method can be adjusted for nonadditive measurement-error (Carroll et al., 2006; Biewen et al., 2008). In particular we are interested in the circumstances of multiplicative error. That is, the case when the observed model design is the true design multiplied by some measurement-error. Using the same notation as before,

$$\mathbf{W}_i = \mathbf{X}_i \odot \mathbf{U}_i$$

where $\odot$ denotes element-wise multiplication and the error is now log-normally distributed with $\log(\mathbf{U}_i) \sim N_p(\mathbf{0}_p, \sigma_u^2 \mathbf{I}_p)$. The log-normal distribution has probability density

$$f(u) = \frac{1}{u\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log(u)-\mu)^2}{2\sigma^2}\right\},$$

where μ is the log-mean and σ^2 is the log-variance. Then for a log-normally distributed random variable, $U \sim \text{lnN}(\mu, \sigma^2)$ the expectation and variance respectively will be,

$$\mathbb{E}(U) = e^{\mu+\frac{1}{2}\sigma^2}, \;\; \text{Var}(U) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2},$$

or for the multivariate case, if $\mathbf{U} \sim \text{lnN}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_u)$,

$$\mathbb{E}(\mathbf{U})_i = e^{\mu_i+\frac{1}{2}\boldsymbol{\Sigma}_{u,ii}}, \;\; \text{Var}(\mathbf{U})_{u,ij} = e^{\mu_i+\mu_j+\frac{1}{2}(\boldsymbol{\Sigma}_{u,ii}+\boldsymbol{\Sigma}_{u,jj})}\left(e^{\boldsymbol{\Sigma}_{u,ij}} - 1\right).$$

This means that the expectation of the measurement-error contaminated design $\mathbf{W}_i$, is $\mathbf{X}_i$ and the errors are scaled according to the size of the original elements. In many cases this is more realistic than simple additive measurement-error as the relative scaling of the design is preserved. In MEG examples this can be reasoned that the forward model for locations closer to a sensor will likely be more affected by error than those for more distant locations.

There are two ways that we can apply the simulation extrapolation method to multiplicative measurement-error.

5.2.1 *Log transformed SIMEX*

In non-linear problems the first instinct is often to find a transformation that will linearise the problem. Indeed many measurement-error models can be transformed to the simple additive case (Eckert et al., 1997). The easiest way to view the multiplicative problem under the SIMEX framework is therefore to linearise the simulation step through a log transform. Under this transformation we simulate new samples of the model design as follows,

$$\mathbf{W}_{i,b} = \exp\{\log(\mathbf{W}_i) + \sqrt{\zeta}\log(\mathbf{U}_{i,b})\}$$

$b = 1, ..., B$. The rest of the SIMEX method is followed as before and the same extrapolant functions can be used as before. The downside of this approach is that in order for us to be able to use the log transform, we require all the values of the covariate design to be positive. This restriction can be overcome by a simple modification of the simulation step as detailed in the next section.

5.2.2 *Multiplicative SIMEX*

In the multiplicative SIMEX approach (Nolte, 2007) we multiply the measurement-error contaminated variables by additional error. Therefore for error level controlling parameter ζ, we simulate B new observations of the covariates,

$$\mathbf{W}_{i,b} = \mathbf{W}_i \odot \mathbf{U}_{i,b}^{\zeta}$$

where each $\mathbf{U}_{i,b}$ is simulated from a log-normal distribution with $\mathbb{E}(\mathbf{U}_{i,b}) = \mathbf{1}_p$ and $\text{Var}(\log(\mathbf{U}_{i,b})) = \sigma_u^2 \mathbf{I}_p$. The averaging and extrapolation steps remain the same as the standard linear SIMEX method. This can be seen by again considering the variance of the simulated measurement-error contaminated matrix, $\mathbf{W}_{i,b}$. Assuming that $\mathbf{U}_i$ has mean $\mathbf{1}_p$, and is independent of $\mathbf{X}_i$,

$$\begin{aligned}
\text{Var}(\mathbf{W}_{i,b}) &= \text{Var}\left(\mathbf{W}_i \odot \mathbf{U}_{i,b}^{\zeta}\right) = \text{Var}\left((\mathbf{X}_i \odot \mathbf{U}_i) \odot \mathbf{U}_{i,b}^{\zeta}\right) = \text{Var}\left(\mathbf{X}_i \odot \mathbf{U}_{i,b}^{\zeta+1}\right) \\
&= \mathbb{E}(\mathbf{X}_i)^2 \text{Var}\left(\mathbf{U}_{i,b}^{\zeta+1}\right) + \text{Var}(\mathbf{X}_i)\mathbb{E}\left(\mathbf{U}_{i,b}^{\zeta+1}\right)^2 + \text{Var}(\mathbf{X}_i)\text{Var}\left(\mathbf{U}_{i,b}^{\zeta+1}\right) \\
&= \text{Var}\left(\mathbf{U}_{i,b}^{\zeta+1}\right)\left(\mathbb{E}(\mathbf{X}_i)^2 + \text{Var}(\mathbf{X}_i)\right) + \text{Var}(\mathbf{X}_i) \\
&= \mathbb{E}\left(\left(\mathbf{U}_{i,b}^{\zeta+1}\right)^2\right)\mathbb{E}(\mathbf{X}_i^2) - \mathbb{E}\left(\mathbf{U}_{i,b}^{\zeta+1}\right)^2 \mathbb{E}(\mathbf{X}_i^2) + \text{Var}(\mathbf{X}_i) \\
&= \mathbb{E}\left(\left(\mathbf{U}_{i,b}^{\zeta+1}\right)^2\right)\mathbb{E}(\mathbf{X}_i^2) - \mathbb{E}(\mathbf{X}_i)^2.
\end{aligned}$$

Hence, the extrapolation $\zeta = -1$ is again required for $\text{Var}(\mathbf{W}_{i,b}) = \text{Var}(\mathbf{X}_i)$.

5.2.3 *Simulations*

In order to look at the effect of multiplicative measurement-error on sparse methods and apply the multiplicative SIMEX we perform a similar simulation to the additive error from chapter 3. For each of 250 Monte Carlo repetitions, 100 observations are simulated from 500 parameters (5 of which are non-zero and sampled from $N(0, 2^2)$), a design matrix $\mathbf{X}_i \sim N(\mathbf{0}, \mathbf{I})$, and additive noise $\epsilon_i \sim N(0, 0.1)$. The measurement-error is then simulated from a log-normal distribution with expectation 1, i.e. $\mathbf{U}_i \sim \text{lnN}(-\frac{\sigma_u^2}{2} \times \mathbf{1}, \sigma_u^2 \mathbf{I})$. Four levels of measurement-error are used, $\sigma_u^2 \in \{0.01, 0.04, 0.16, 0.25\}$, and for each level the design matrix is component wise multiplied by the simulated measurement-error matrix in order to give the error contaminated design. The naive estimates for each of the methods (min-norm, lasso, elastic net, square root lasso and PED) are then computed with the measurement-error contaminated design matrix. Since the asymptotic penalty in the square root lasso produced estimates that were too sparse in the additive simulations, BIC was used to choose the penalty level instead. For each method the estimate under no measurement-error is also calculated and the estimates for each error level are assessed for bias, the number of true positives and false positives, as well as the proportion of signal detected at the true positives. The results for the naive estimates are found in Table 10.

Method	σ_u^2	Bias	True +	False +	Prop-tr
Ridge	0	18.58 (6.73)	5 (0)	495 (0)	0.07 (0.01)
	0.01	18.61 (6.67)	5 (0)	495 (0)	0.07 (0.01)
	0.04	18.56 (6.77)	5 (0)	495 (0)	0.07 (0.01)
	0.16	18.25 (6.77)	5 (0)	495 (0)	0.07 (0.01)
	0.25	17.71 (6.75)	5 (0)	495 (0)	0.06 (0.02)
Lasso	0	0.78 (0.27)	4.83 (0.43)	25.52 (13.86)	0.95 (0.03)
	0.01	1.33 (0.53)	4.76 (0.51)	25.59 (15.37)	0.91 (0.03)
	0.04	2.32 (1.22)	4.57 (0.66)	24.56 (16.23)	0.83 (0.05)
	0.16	4.47 (2.13)	4.20 (0.82)	23.87 (17.63)	0.65 (0.07)
	0.25	5.41 (2.68)	4.12 (0.82)	24.30 (19.69)	0.57 (0.08)
E-net	0	1.18 (0.34)	4.84 (0.43)	37.99 (13.53)	0.93 (0.03)
	0.01	1.96 (0.69)	4.75 (0.51)	38.34 (14.73)	0.88 (0.03)
	0.04	3.33 (1.57)	4.58 (0.64)	39.15 (18.60)	0.80 (0.05)
	0.16	5.72 (2.62)	4.26 (0.80)	35.77 (19.79)	0.61 (0.07)
	0.25	6.52 (3.10)	4.13 (0.85)	33.50 (21.55)	0.52 (0.08)
SRL	0	0.76 (0.58)	4.81 (0.45)	16.38 (3.88)	0.93 (0.04)
	0.01	1.24 (0.73)	4.74 (0.53)	15.96 (3.98)	0.89 (0.04)
	0.04	2.14 (1.07)	4.56 (0.65)	16.84 (4.57)	0.81 (0.05)
	0.16	4.09 (1.78)	4.24 (0.78)	16.85 (4.11)	0.62 (0.07)
	0.25	4.86 (2.03)	4.06 (0.83)	16.90 (4.55)	0.54 (0.08)
PED	0	0.81 (0.98)	4.23 (0.75)	6.08 (8.52)	0.93 (0.06)
	0.01	1.53 (1.34)	4.14 (0.79)	8.99 (9.36)	0.89 (0.07)
	0.04	3.57 (1.94)	4.00 (0.85)	16.07 (8.34)	0.79 (0.09)
	0.16	9.15 (3.72)	3.85 (0.90)	27.72 (5.95)	0.56 (0.09)
	0.25	11.29 (4.23)	3.78 (0.90)	32.30 (5.90)	0.48 (0.09)

Table 10: Naive estimates: bias, true positives, false positives and proportion of signal at true positives. Standard deviation in parenthesis. 250 Monte Carlo simulations.

The ridge (minimum norm) method is unsuitable for the level of sparsity displayed in the data and as such performs badly in terms of both bias, false positives and signal at the true positives. The sparse methods perform well in the lower levels of measurement-error, with the lasso giving the best performance due to the uncorrelated nature of the simulated covariates. As we increase the measurement-error the bias increases, and the mean number of true positives, as well as the signal at those true positives, reduces.

The introduction of measurement-error is especially detrimental to the performance of the PED estimate. For the lower levels of measurement-error, PED performs quite well with respect to false positives. Surprisingly, it even performs better than the square root lasso in this regard, although this is probably due to the chosen threshold level. However, the number of false positives and the overall bias both increase significantly as we introduce higher levels of measurement-error to the PED estimate.

We now perform the multiplicative SIMEX method as detailed in section 5.2.2 for each of the methods in Table 10. For each Monte Carlo repetition from the original simulation we have $B = 50$ repeated samples of $\mathbf{U}_{i,b}$ in the simulation step and estimate for $\zeta \in \{0.5, 1, 1.5, 2\}$. A quadratic model is then employed in the extrapolation step in order to give the SIMEX estimate. Since the SIMEX method requires zeroes at each repetition and level of the simulation step in order to give a zero estimate in the extrapolation, we often get a large number of small, but otherwise non-zero parameter estimates. Therefore thresholding was employed in order to remove some of the smallest estimates from the model. This was achieved by employing the same approach as for PED (Vasiliu et al., 2014) and setting to zero the $\hat{\beta}_j$'s where $\frac{|\hat{\beta}_j|}{||\hat{\boldsymbol{\beta}}||_2} < \frac{C}{\sqrt{n}}$, with $C = 0.2$.

Method	σ_u^2	Bias	True +	False +	Prop-tr
Ridge	0.01	42.84 (15.81)	5 (0)	495 (0)	0.22 (0.04)
	0.04	42.66 (15.46)	5 (0)	495 (0)	0.22 (0.04)
	0.16	41.31 (15.93)	5 (0)	495 (0)	0.20 (0.04)
	0.25	39.29 (15.68)	5 (0)	495 (0)	0.18 (0.05)
Lasso	0.01	1.35 (0.72)	4.66 (0.57)	4.92 (6.32)	0.89 (0.04)
	0.04	2.93 (2.06)	4.51 (0.68)	9.74 (9.04)	0.80 (0.06)
	0.16	5.98 (3.67)	4.18 (0.81)	19.60 (10.68)	0.69 (0.10)
	0.25	7.10 (5.32)	4.03 (0.84)	21.20 (12.64)	0.70 (0.12)
E-net	0.01	1.74 (0.77)	4.68 (0.56)	7.07 (6.40)	0.87 (0.04)
	0.04	3.89 (2.33)	4.51 (0.68)	15.89 (11.22)	0.78 (0.06)
	0.16	7.57 (4.53)	4.17 (0.82)	27.92 (12.41)	0.66 (0.10)
	0.25	8.53 (5.50)	4.04 (0.85)	28.87 (13.53)	0.64 (0.13)
SRL	0.01	1.16 (0.68)	4.66 (0.56)	1.90 (1.58)	0.87 (0.04)
	0.04	2.25 (1.01)	4.49 (0.67)	5.14 (2.49)	0.79 (0.05)
	0.16	4.20 (1.88)	4.16 (0.78)	10.81 (3.37)	0.67 (0.08)
	0.25	4.88 (2.10)	3.95 (0.86)	13.37 (3.80)	0.65 (0.11)
PED	0.01	2.14 (1.67)	4.19 (0.77)	6.35 (6.49)	0.87 (0.10)
	0.04	4.96 (2.35)	4.10 (0.85)	19.05 (7.08)	0.79 (0.11)
	0.16	15.75 (6.39)	4.06 (0.85)	56.69 (8.11)	0.65 (0.14)
	0.25	20.87 (8.05)	4.05 (0.83)	68.66 (7.98)	0.59 (0.15)

Table 11: SIMEX estimates with sparse thresholding: bias, true positives, false positives and proportion of signal at true positives. Standard deviation in parenthesis. 250 Monte Carlo simulations.

For the lower measurement-error levels, the SIMEX estimates offer little improvement over the naive estimates. In fact the increase in bias suggests that

SIMEX performs worse than the naive methods. However for the larger error levels the SIMEX estimates give an improvement in the estimates at the true positives. The increase in bias also suggests that the false positives that have not been removed with the thresholding also increase in magnitude. Fig. 32 represents the true values, naive estimates and SIMEX corrections for the lasso from a single Monte Carlo repetition at each of the four measurement-error levels.

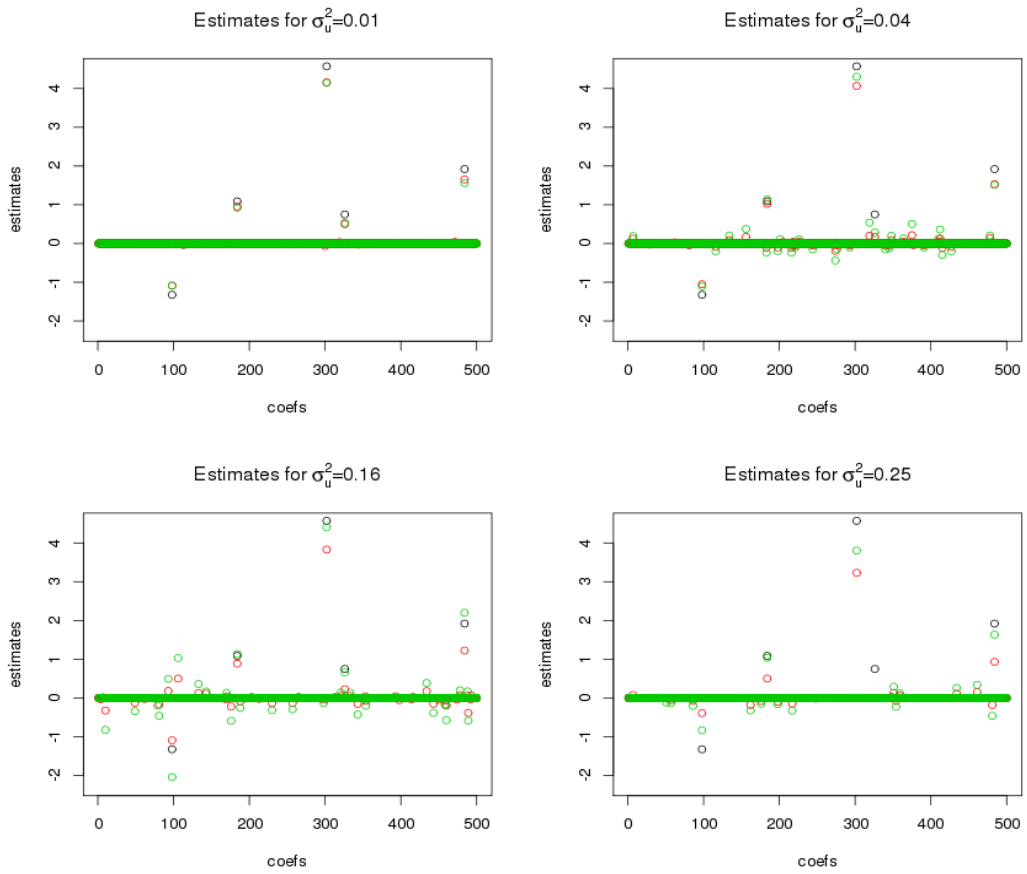

Figure 32: Naive (red) and SIMEX (green) estimates for the lasso when, $\sigma_u^2 = 0.01, 0.04, 0.16$ and 0.25. True values in black.

We can see from the plots in Fig. 32 that both the true and false positives are increased in the SIMEX estimate. However it is also evident that the SIMEX correction gives a greater improvement over the naive estimate under larger

values of measurement-error. Despite this, as we see in the plot for $\sigma_u^2 = 0.25$ the SIMEX method is still dependent on the selection abilities of the naive estimate and it appears from the empirical evidence that SIMEX is unable to correct for covariates that are erroneously excluded.

In conclusion the simulation extrapolation method in general gives an improvement over the naive estimates in both the additive and multiplicative measurement-error settings. However it is generally restricted to rescaling correction and is unable to correct poor selection. The basis of SIMEX in simulation means that it is naturally computationally expensive. When we include the naive estimate, SIMEX requires us to produce $B \times n_\zeta + 1$ estimates, where n_ζ is the number of levels of ζ used. In addition to this we have to find an average solution for each level and then extrapolate the SIMEX estimate. As we scale up the dimensions of our data, SIMEX will become slower and will require more computation to employ. For the large number of locations involved in the MEG problem it may be worth reducing the dimensions before SIMEX is used. The application of SIMEX to real data will be studied in chapter 8.

6

THE CORRECTED ELASTIC NET

So far two methods have been considered for the correction of measurement-error in sparse regression methods. The conditional score method involves solving score functions that include a sufficient statistic for the error free design matrix, and in SIMEX we use the estimation performance under increasing error to extrapolate an estimate for the error free case. Loh and Wainwright (2012) proposed a simple correction method for the lasso to help deal with noisy or missing data. The corrected lasso takes the expected value of the squared loss in the presence of measurement-error as its starting point, from which a correction is determined. Further work on the corrected lasso has been done by Sørensen et al. (2015) who investigated the impact of measurement-error on the naive lasso and the circumstances under which the naive estimate gives good results. In this chapter we extend this work to cover the elastic net. This corrected elastic net has advantages over the measurement-error methods previously covered in that it is specifically developed for sparse methods, where the number of covariates may be large.

6.1 CORRECTED LASSO

The corrected lasso can be arrived at by assessing the performance of the standard (naive) lasso in the presence of measurement-error. We begin by considering the expected loss function under measurement-error,

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n}||\mathbf{y}-\mathbf{W}\boldsymbol{\beta}||_2^2|\mathbf{X},\mathbf{y}\right) &= \frac{1}{n}\mathbb{E}\left(||\mathbf{y}-(\mathbf{X}+\mathbf{U})\boldsymbol{\beta}||_2^2\right)\\ &= \frac{1}{n}\Big(\mathbf{y}^T\mathbf{y}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}-2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta}\\ &\quad +\mathbb{E}\left(\boldsymbol{\beta}^T\mathbf{U}^T\mathbf{U}\boldsymbol{\beta}+2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{U}\boldsymbol{\beta}-2\mathbf{y}^T\mathbf{U}\boldsymbol{\beta}\right)\Big)\\ &= \frac{1}{n}||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||_2^2+\boldsymbol{\beta}^T\boldsymbol{\Sigma}_u\boldsymbol{\beta}.\end{aligned}$$

The expected loss function is therefore biased. This motivates a measurement-error corrected version of the lasso

$$\boldsymbol{\beta}_{CL}=\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\frac{1}{n}||\mathbf{y}-\mathbf{W}\boldsymbol{\beta}||_2^2-\boldsymbol{\beta}^T\boldsymbol{\Sigma}_u\boldsymbol{\beta}+\lambda||\boldsymbol{\beta}||_1\right\}.$$

We can alternatively view this problem as an M-estimator of the form

$$\boldsymbol{\beta}_{CL}=\arg\min_{\boldsymbol{\beta}:||\boldsymbol{\beta}||_1\leq\kappa}\left\{\boldsymbol{\beta}^T\left(\frac{1}{n}\mathbf{W}^T\mathbf{W}-\boldsymbol{\Sigma}_u\right)\boldsymbol{\beta}-\frac{2}{n}\mathbf{y}^T\mathbf{W}\boldsymbol{\beta}\right\}.$$

6.2 PERFORMANCE OF ELASTIC NET UNDER MEASUREMENT-ERROR

We now investigate the performance of the elastic net under additive measurement-error and extend the idea of the corrected lasso to incorporate an additional ℓ_2 norm term. The following theoretical work is generalised for the elastic net. The corresponding results for the lasso as previously given by Sørensen et al. (2015) can be arrived at by simply setting $\lambda_2=0$. Let $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}^0$, where $\boldsymbol{\beta}^0$ is

the vector of the true covariate values. In the presence of measurement-error, the naive solution of the elastic net is defined as

$$\hat{\boldsymbol{\beta}}_{Naive-EN} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{n}||\mathbf{y} - \mathbf{W}\boldsymbol{\beta}||_2^2 + \lambda_2||\boldsymbol{\beta}||_2^2 + \lambda_1||\boldsymbol{\beta}||_1 \right\}.$$

Define the active set of the elastic net estimate for a chosen λ_1 as $S(\lambda_1) = \{j : \hat{\beta}_j \neq 0\}$. Given $\boldsymbol{\beta}_0$, we then order the covariates according to relevant and non-relevant covariates, i.e. $S_0 = \{1, ..., s_0\}$, $S_0^c = \{s_0 + 1, ..., p\}$ where $s_0 << p$ is the cardinality of S.

We use these orderings to partition the matrices $\mathbf{W}, \mathbf{X}$ and $\mathbf{U}$, such as $\mathbf{W} = \left(\mathbf{W}_{S_0}, \mathbf{W}_{S_0^c}\right)$. We then denote covariances by $\mathbf{C}$ where the subscripts show which covariates are involved. Using $\mathbf{W}$ as an example, $\mathbf{C}_{ww} = \frac{1}{n}\mathbf{W}^T\mathbf{W}$. The covariance matrix can also be partitioned according to the covariate relevance as follows,

$$\mathbf{C}_{ww} = \begin{pmatrix} \mathbf{C}_{ww}(S_0, S_0) & \mathbf{C}_{ww}(S_0, S_0^c) \\ \mathbf{C}_{ww}(S_0^c, S_0) & \mathbf{C}_{ww}(S_0^c, S_0^c) \end{pmatrix}.$$

Population covariance matrices follow the same indexing but are denoted by $\boldsymbol{\Sigma}$.

6.2.1 *Selection*

We now assess the ability of the elastic net for recovering the correct signs of the covariates. For the following section we have adapted the work of Sørensen et al. (2015) in an obvious way, extending their results to the elastic net. In practice this involves replacing $\mathbf{C}_{w,w}$ and $\mathbf{C}_{w,u}$ with $(\mathbf{C}_{w,w} + \lambda_2\mathbf{I})$ and $(\mathbf{C}_{w,u} + \lambda_2\mathbf{I})$ respectively. For completeness we repeat Sørensen et al.'s results with the changes for the elastic net implemented.

Selection and sign consistency are important aspects of sparse modelling for high dimensional problems. In cases without measurement-error the lasso re-

quires an Irrepresentable Condition (IC) in order to provide sign consistent selection (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006). The Irrepresentable Condition was extended for elastic net models with the Elastic Irrepresentable Condition (EIC) (Jia and Yu, 2010). In the measurement-error model, the EIC is,

Definition 6.2.1. Elastic Irrepresentable Condition with measurement-error (EIC-ME). There exists a constant $\theta \in [0, 1)$ such that,

$$\left\| \mathbf{C}_{ww}(S_0^c, S_0) \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \left(\text{sign}(\boldsymbol{\beta}^0_{S_0}) + 2\frac{\lambda_2}{\lambda_1} \boldsymbol{\beta}^0_{S_0} \right) \right\|_\infty \leq \theta.$$

As with the lasso, in the presence of measurement-error we additionally require a Measurement-Error Condition in order to give a high probability of sign consistency (Sørensen et al., 2015).

Definition 6.2.2. Measurement-Error Condition (MEC). The MEC is satisfied if

$$\boldsymbol{\Sigma}_{ww}(S_0^c, S_0)\boldsymbol{\Sigma}_{ww}(S_0, S_0)^{-1}\boldsymbol{\Sigma}_{uu}(S_0, S_0) - \boldsymbol{\Sigma}_{uu}(S_0^c, S_0) = \mathbf{0}.$$

The MEC is a condition on the population covariance matrices rather than sample covariances. In the following result, the EIC-ME is used to bound the probability of sign consistent selection for the naive elastic net. The MEC is a sufficient condition which enables the probability of sign consistency to asymptotically approach 1.

Lemma 1. If we assume the EIC-ME holds for constant θ, then

$$\mathbb{P}\left(\text{sign}\left(\hat{\boldsymbol{\beta}}\right) = \text{sign}\left(\boldsymbol{\beta}^0\right)\right) \geq \mathbb{P}(A \cap B)$$

where the events A and B are defined as,

$$A = \left\{ |\mathbf{Z}_1\boldsymbol{\epsilon} - \mathbf{Z}_2\boldsymbol{\beta}^0_{S_0}| < \sqrt{n}\left(|\boldsymbol{\beta}^0_{S_0}| - \frac{\lambda_1}{2} \left| \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \left(\text{sign}(\boldsymbol{\beta}^0_{S_0}) \right) \right| \right) \right\},$$

$$B = \left\{ |\mathbf{Z}_3\boldsymbol{\epsilon} - \mathbf{Z}_4\boldsymbol{\beta}^0_{S_0}| < \frac{\lambda_1}{2}(1-\theta)\mathbf{1} \right\},$$

where,

$$\mathbf{Z}_1 = \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I}\right)^{-1} \frac{\mathbf{W}^T_{S_0}}{\sqrt{n}}$$

$$\mathbf{Z}_2 = \sqrt{n}\left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I}\right)^{-1}\left(\mathbf{C}_{wu}(S_0, S_0) + \lambda_2\mathbf{I}\right)$$

$$\mathbf{Z}_3 = \mathbf{C}_{ww}(S_0^c, S_0)\left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I}\right)^{-1}\frac{\mathbf{W}^T_{S_0}}{\sqrt{n}} - \frac{\mathbf{W}^T_{S_0^c}}{\sqrt{n}}$$

$$\mathbf{Z}_4 = \sqrt{n}\left(\mathbf{C}_{ww}(S_0^c, S_0)\left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I}\right)^{-1}\mathbf{C}_{wu}(S_0, S_0) - \mathbf{C}_{wu}(S_0^c, S_0)\right).$$

Proof. We start by deriving the Karush-Kuhn-Tucker (KKT) conditions (Karush, 1939; Kuhn and Tucker, 1951). Introduce a new coefficient $\boldsymbol{\gamma} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$, then the naive elastic net under measurement-error is equivalent to

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}}\left(-\frac{2}{n}\boldsymbol{\epsilon}^T\mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\gamma}^T(\mathbf{C}_{ww} + \lambda_2\mathbf{I})\boldsymbol{\gamma} + 2\boldsymbol{\gamma}^T(\mathbf{C}_{wu} + \lambda_2\mathbf{I})\boldsymbol{\beta}^0 + \lambda_1||\boldsymbol{\gamma} + \boldsymbol{\beta}^0||_1\right)$$

where we include only those terms that depend on $\boldsymbol{\gamma}$. We arrive at the KKT conditions by differentiating with respect to $\boldsymbol{\gamma}$.

Therefore by the KKT conditions, $\hat{\boldsymbol{\gamma}}$ is an optimal solution if and only if,

$$-\frac{2}{n}\boldsymbol{\epsilon}^T\mathbf{W} + 2(\mathbf{C}_{ww} + \lambda_2\mathbf{I})\hat{\boldsymbol{\gamma}} + 2(\mathbf{C}_{wu} + \lambda_2\mathbf{I})\boldsymbol{\beta}^0 + \lambda_1\hat{\boldsymbol{\tau}} = 0,$$

where

$$\hat{\tau}_j = \begin{cases} \operatorname{sign}(\hat{\beta}_j), & \hat{\beta}_j \neq 0 \\ \text{Real number} \in [-1, 1], & \hat{\beta}_j = 0. \end{cases}$$

The rest of the proof then follows by considering the KKT conditions under the partitions defined earlier. The structure follows that of Zhao and Yu (2006) and Sørensen et al. (2015) who proved the similar results for the lasso and the lasso under measurement-error respectively.

By the KKT conditions, the naive elastic net will give sign consistent selection, i.e. $\text{sign}(\hat{\boldsymbol{\beta}}_{S_0}) = \text{sign}(\boldsymbol{\beta}^0_{S_0})$ and $\hat{\boldsymbol{\beta}}_{S_0^c} = \mathbf{0}$, if there exists a $\hat{\boldsymbol{\gamma}}$ and the following hold,

$$-\frac{\mathbf{W}^T_{S_0}}{\sqrt{n}}\boldsymbol{\epsilon} + \sqrt{n}\left((\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I})\hat{\boldsymbol{\gamma}}_{S_0} + (\mathbf{C}_{wu}(S_0, S_0) + \lambda_2\mathbf{I})\boldsymbol{\beta}^0_{S_0}\right) = -\frac{\lambda_1\sqrt{n}}{2}\text{sign}(\boldsymbol{\beta}^0_{S_0}) \tag{6.1}$$

$$|\hat{\boldsymbol{\gamma}}_{S_0}| < |\boldsymbol{\beta}^0_{S_0}| \tag{6.2}$$

$$\left|-\frac{\mathbf{W}^T_{S_0^c}}{\sqrt{n}}\boldsymbol{\epsilon} + \sqrt{n}\mathbf{C}_{ww}(S_0^c, S_0)\hat{\boldsymbol{\gamma}}_{S_0} + \sqrt{n}\mathbf{C}_{wu}(S_0^c, S_0)\boldsymbol{\beta}^0_{S_0}\right| \leq \frac{\lambda_1\sqrt{n}}{2}\mathbf{1}. \tag{6.3}$$

Note, in the third condition the λ_2 terms drop from the inequality as the partition (S_0^c, S_0) when applied to the identity matrix is simply a matrix of zeroes. We show that if the events A and B hold then the three KKT conditions above are satisfied. Starting with event A,

$$|\mathbf{Z}_1\boldsymbol{\epsilon} - \mathbf{Z}_2\boldsymbol{\beta}^0_{S_0}| < \sqrt{n}\left(|\boldsymbol{\beta}^0_{S_0}| - \frac{\lambda_1}{2}\left|(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I})^{-1}\left(\text{sign}(\boldsymbol{\beta}^0_{S_0})\right)\right|\right),$$

we can say that there exists $|\hat{\boldsymbol{\gamma}}_{S_0}| < |\boldsymbol{\beta}^0_{S_0}|$ so that

$$|\mathbf{Z}_1\boldsymbol{\epsilon} - \mathbf{Z}_2\boldsymbol{\beta}^0_{S_0}| = \sqrt{n}\left(|\hat{\boldsymbol{\gamma}}_{S_0}| - \frac{\lambda_1}{2}\left|(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I})^{-1}\left(\text{sign}(\boldsymbol{\beta}^0_{S_0})\right)\right|\right).$$

There also must exist $|\hat{\boldsymbol{\gamma}}_{S_0}| < |\boldsymbol{\beta}^0_{S_0}|$ such that

$$\mathbf{Z}_1\boldsymbol{\epsilon} - \mathbf{Z}_2\boldsymbol{\beta}^0_{S_0} = \sqrt{n}\left(\hat{\boldsymbol{\gamma}}_{S_0} - \frac{\lambda_1}{2}\left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I}\right)^{-1}\left(\text{sign}(\boldsymbol{\beta}^0_{S_0})\right)\right),$$

choosing the appropriate sign for the elements of $\hat{\boldsymbol{\gamma}}_{S_0}$. We can also have $|\hat{\boldsymbol{\gamma}}_{S_0}| < |\boldsymbol{\beta}^0_{S_0}|$ such that

$$-(\mathbf{Z}_1\boldsymbol{\epsilon} - \mathbf{Z}_2\boldsymbol{\beta}^0_{S_0}) = \sqrt{n}\left(-\hat{\boldsymbol{\gamma}}_{S_0} - \frac{\lambda_1}{2}\left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I}\right)^{-1}\left(\text{sign}(\boldsymbol{\beta}^0_{S_0})\right)\right).$$

Multiplying this by $(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I})$ and rearranging the terms we get to the first condition. Hence, event A satisfies the first two conditions. Now we add and subtract $\sqrt{n}\mathbf{C}_{ww}(S_0^c, S_0)\hat{\boldsymbol{\gamma}}_{S_0}$ from the LHS of B and use the triangle inequality to give,

$$\left|\left(\mathbf{C}_{ww}(S_0^c,S_0)\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\frac{\mathbf{W}_{S_0}^T}{\sqrt{n}}-\frac{\mathbf{W}_{S_0^c}^T}{\sqrt{n}}\right)\boldsymbol{\epsilon}+\sqrt{n}\mathbf{C}_{ww}(S_0^c,S_0)\hat{\boldsymbol{\gamma}}_{S_0}\right.$$
$$-\sqrt{n}\left(\mathbf{C}_{ww}(S_0^c,S_0)\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\mathbf{C}_{wu}(S_0,S_0)-\mathbf{C}_{wu}(S_0^c,S_0)\right)\boldsymbol{\beta}_{S_0}^0$$
$$\left.-\sqrt{n}\mathbf{C}_{ww}(S_0^c,S_0)\hat{\boldsymbol{\gamma}}_{S_0}\right|<\frac{\lambda_1}{2}(1-\theta)\mathbf{1}$$

which implies,

$$\left|-\frac{\mathbf{W}_{S_0^c}^T}{\sqrt{n}}\boldsymbol{\epsilon}+\sqrt{n}\mathbf{C}_{ww}(S_0^c,S_0)\hat{\boldsymbol{\gamma}}_{S_0}+\sqrt{n}\mathbf{C}_{wu}(S_0^c,S_0)\boldsymbol{\beta}_{S_0}^0\right|-$$
$$\left|-\mathbf{C}_{ww}(S_0^c,S_0)\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\frac{\mathbf{W}_{S_0}^T}{\sqrt{n}}\boldsymbol{\epsilon}+\sqrt{n}\mathbf{C}_{ww}(S_0^c,S_0)\hat{\boldsymbol{\gamma}}_{S_0}\right.$$
$$\left.+\sqrt{n}\mathbf{C}_{ww}(S_0^c,S_0)\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\mathbf{C}_{wu}(S_0,S_0)\boldsymbol{\beta}_{S_0}^0\right|\leq\frac{\lambda_1}{2}(1-\theta)\mathbf{1}\,.$$

Rearranging the first condition gives,

$$\begin{aligned}\sqrt{n}\hat{\boldsymbol{\gamma}}=&(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I})^{-1}\frac{\mathbf{W}_{S_0}^T}{\sqrt{n}}\boldsymbol{\epsilon}-\sqrt{n}(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I})^{-1}\mathbf{C}_{wu}(S_0,S_0)\boldsymbol{\beta}_{S_0}^0\\&-\sqrt{n}\frac{\lambda_1}{2}\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\left(\operatorname{sign}(\boldsymbol{\beta}_{S_0}^0)+2\frac{\lambda_2}{\lambda_1}\boldsymbol{\beta}_{S_0}^0\right).\end{aligned}$$

Now substituting this into the second term on the LHS we have

$$\left|-\frac{\mathbf{W}_{S_0^c}^T}{\sqrt{n}}\boldsymbol{\epsilon}+\sqrt{n}\mathbf{C}_{ww}(S_0^c,S_0)\hat{\boldsymbol{\gamma}}_{S_0}+\sqrt{n}\mathbf{C}_{wu}(S_0^c,S_0)\boldsymbol{\beta}_{S_0}^0\right|-$$
$$\left|\frac{\lambda_1\sqrt{n}}{2}\mathbf{C}_{ww}(S_0^c,S_0)\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\left(\operatorname{sign}(\boldsymbol{\beta}_{S_0}^0)+2\frac{\lambda_2}{\lambda_1}\boldsymbol{\beta}_{S_0}^0\right)\right|\leq\frac{\lambda_1\sqrt{n}}{2}(1-\theta).$$

The second term on the LHS is now equal to $\frac{\lambda\sqrt{n}}{2}$ times the EIC-ME condition. Hence, adding $\frac{\lambda\sqrt{n}}{2}\theta$ to both sides we have

$$\left|-\frac{\mathbf{W}_{S_0^c}^T}{\sqrt{n}}\boldsymbol{\epsilon}+\sqrt{n}\mathbf{C}_{ww}(S_0^c,S_0)\hat{\boldsymbol{\gamma}}_{S_0}+\sqrt{n}\mathbf{C}_{wu}(S_0^c,S_0)\boldsymbol{\beta}_{S_0}^0\right|\leq\frac{\lambda_1\sqrt{n}}{2}\mathbf{1},$$

which is condition three. Hence given A, B satisfies the third condition.

□

The events A and $B|A$ correspond to the correct sign being chosen on the true coefficients and the non-relevant coefficients being set to zero respectively. Due

to the presence of measurement-error, the left hand sides of the events involve $\boldsymbol{\beta}^0_{S_0}$ terms as well as the noise $\boldsymbol{\epsilon}$. This means that the selection performance is also dependent on the product of the measurement-error and the true values of the covariates. Also implicit in the events is a trade off in the choice of the lambdas (particularly λ_1), such that we select the relevant β_j's but provide a sparsity level that is able to discard irrelevant covariates.

Theorem 1 provides an asymptotic bound on the probability of the events A and B from Lemma 1 being satisfied and hence sign consistency.

Theorem 1. Suppose that the MEC is satisfied and

$$|\boldsymbol{\beta}^0_{S_0}| > |\boldsymbol{\Sigma}_{ww}(S_0, S_0)^{-1}\boldsymbol{\Sigma}_{uu}(S_0, S_0)\boldsymbol{\beta}^0_{S_0}|.$$

We also assume the following regularity conditions; Given fixed p and q, we assume

$$\frac{1}{n}\mathbf{X}^T\mathbf{X} = \mathbf{C}_{xx} \to \boldsymbol{\Sigma}_{xx}, \text{ as } n \to \infty$$

and

$$\max_{1\leq i\leq n}(\mathbf{x}_i^T\mathbf{x}_i)\frac{1}{n} \to 0, \text{ as } n \to \infty.$$

Similarly, $\mathbf{C}_{uu} \to \boldsymbol{\Sigma}_{uu}$, $\mathbf{C}_{ww} \to \boldsymbol{\Sigma}_{ww}$, $\max_{1\leq i\leq n}(\mathbf{u}_i^T\mathbf{u}_i)\frac{1}{n} \to 0$ and $\max_{1\leq i\leq n}(\mathbf{w}_i^T\mathbf{w}_i)\frac{1}{n} \to 0$ as $n \to \infty$. Finally $\sqrt{n}(\mathbf{C}_{uu} - \boldsymbol{\Sigma}_{uu})$ and $\sqrt{n}(\mathbf{C}_{ww} - \boldsymbol{\Sigma}_{ww})$ converge to normal distributions with mean zero and finite covariances. Therefore we have assumed that the sample covariance matrices will approach the population covariances and that the central limit theorem is appropriate. Similar conditions have been assumed by Knight and Fu (2000) and Zhao and Yu (2006).

Now if λ_1,$\lambda_2 \to 0$ and $\lambda_1 n^{(1-c)/2} \to \infty$ as $n \to \infty$ for some $c \in [0, 1)$, then

$$\mathbb{P}\left(\text{sign}(\hat{\boldsymbol{\beta}}) = \text{sign}(\boldsymbol{\beta}^0)\right) = 1 - o(e^{-n^c}).$$

Proof.

$$1 - \mathbb{P}(A \cap B) \leq \mathbb{P}(A^c) + \mathbb{P}(B^c) \leq$$
$$\sum_{j \in S_0} \mathbb{P}\left(\left|(\mathbf{Z}_1\boldsymbol{\epsilon})_j\right| \geq \sqrt{n}\left(\mathbf{a}_j - \frac{\lambda_1}{2}\mathbf{b}_j\right)\right) + \sum_{j \in S_0^c} \mathbb{P}\left(\left|(\mathbf{Z}_3\boldsymbol{\epsilon})_j - \left(\mathbf{Z}_4\boldsymbol{\beta}^0_{S_0}\right)_j\right| \geq \frac{\lambda_1\sqrt{n}}{2}(1-\theta)\right)$$

where,

$$\mathbf{a} = |\boldsymbol{\beta}^0_{S_0}| - \left|\left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I}\right)^{-1}\left(\mathbf{C}_{wu}(S_0, S_0) + \lambda_2\mathbf{I}\right)\boldsymbol{\beta}^0_{S_0}\right| \tag{6.4}$$

$$\mathbf{b} = \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I}\right)^{-1}\left(\text{sign}(\boldsymbol{\beta}^0_{S_0})\right) \tag{6.5}$$

Clearly, $\mathbf{Z}_1\boldsymbol{\epsilon}$ converges in distribution to a normal distribution with mean $\mathbf{0}$ and covariance $\sigma^2\left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I}\right)^{-1}\mathbf{C}_{ww}(S_0, S_0)\left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2\mathbf{I}\right)^{-1}$. Therefore, for $j \in S_0$, there exists a finite κ such that $\mathbb{E}((\mathbf{Z}_1\boldsymbol{\epsilon})_j)^2 < \kappa^2$.

If our assumptions about λ_2 and the regularity conditions hold,

$$\mathbf{a} \to |\boldsymbol{\beta}^0_{S_0}| - \left|\boldsymbol{\Sigma}_{ww}(S_0, S_0)^{-1}\boldsymbol{\Sigma}_{uu}(S_0, S_0)\boldsymbol{\beta}^0_{S_0}\right|,$$

and

$$\mathbf{b} \to \boldsymbol{\Sigma}_{ww}(S_0, S_0)^{-1}\text{sign}(\boldsymbol{\beta}^0), \text{ as } n \to \infty.$$

Then, assuming $\lambda_1 = o(1)$ and using $1 - \Phi(t) < t^{-1}e^{\frac{-t^2}{2}}$,

$$\begin{aligned}
\mathbb{P}(A^c) &\leq \sum_{j \in S_0}\left(1 - \mathbb{P}\left(\frac{|(\mathbf{Z}_1\boldsymbol{\epsilon})_j|}{\kappa} < \frac{\sqrt{n}\mathbf{a}_j}{2\kappa}(1 + o(1))\right)\right) \\
&\leq (1 + o(1))\sum_{j \in S_0}\left(1 - \Phi\left(\frac{\sqrt{n}\mathbf{a}_j}{2\kappa}(1 + o(1))\right)\right) \\
&= o(e^{-n^c}).
\end{aligned}$$

Now moving onto B^c, $\mathbf{Z}_3\boldsymbol{\epsilon}$ converges in distribution to a normal distribution with mean $\mathbf{0}$ and covariance,

$$\sigma^2(\boldsymbol{\Sigma}_{ww}(S_0^c, S_0^c) - \boldsymbol{\Sigma}_{ww}(S_0^c, S_0)\boldsymbol{\Sigma}_{ww}(S_0, S_0)^{-1}\boldsymbol{\Sigma}_{ww}(S_0, S_0^c)).$$

Using the fact that $\sqrt{n}\mathbf{C}_{wu}$ is normally distributed with mean $\sqrt{n}\boldsymbol{\Sigma}_{wu} = \sqrt{n}\boldsymbol{\Sigma}_{uu}$ (Anderson, 2003),

$$\mathbf{Z}_4\boldsymbol{\beta}^0_{S_0} \to \sqrt{n}(\boldsymbol{\Sigma}_{ww}(S_0^c, S_0)\boldsymbol{\Sigma}_{ww}(S_0, S_0)^{-1}\boldsymbol{\Sigma}_{wu}(S_0, S_0) - \boldsymbol{\Sigma}_{wu}(S_0^c, S_0))\boldsymbol{\beta}^0_{S_0}$$

as $n \to \infty$. However, if the MEC holds then this term is equal to zero and hence $\mathbf{Z}_4\boldsymbol{\beta}^0_{S_0}$ is normally distributed with finite covariance.

We can therefore say that there exists κ such that $\mathbb{E}((\mathbf{Z}_3\boldsymbol{\epsilon})_j - (\mathbf{Z}_4\boldsymbol{\beta}^0_{S_0})_j)^2 < \kappa^2$ for $j \in S_0^c$. Then, so long as $\lambda_1 n^{(1-c)/2} \to \infty$ for $0 \leq c < 1$,

$$\begin{aligned}
\mathbb{P}(B^c) &\leq \sum_{j \in S_0^c} \left(1 - \mathbb{P}\left(\frac{|(\mathbf{Z}_3\boldsymbol{\epsilon})_j - (\mathbf{Z}_4\boldsymbol{\beta}^0_{S_0})_j|}{\kappa} < \frac{\sqrt{n}\lambda_1}{2\kappa}(1-\theta)\right)\right) \\
&\leq (1 + o(1)) \sum_{j \in S_0^c} \left(1 - \Phi\left(\frac{\sqrt{n}\lambda_1}{2\kappa}(1-\theta)\right)\right) \\
&= o\left(e^{-n^c}\right).
\end{aligned}$$

From these bounds on the events A^c, B^c, using Lemma 1 we can conclude that the probability of sign consistent selection for the naive elastic net when all stated assumptions hold is,

$$\mathbb{P}(A \cap B) = 1 - o\left(e^{-n^c}\right).$$

□

Therefore the MEC implies that the elastic net is asymptotically sign consistent. We note that the MEC holds if $\boldsymbol{\Sigma}_{xx}(S_0^c, S_0) = \boldsymbol{\Sigma}_{uu}(S_0^c, S_0) = 0$. This corresponds to the relevant and non-relevant covariates being uncorrelated. In reality this is unlikely to be true in most situations. Alternatively, the MEC will also hold if the measurement-error covariance is of the same form as the covariance of the true covariates, i.e. $\boldsymbol{\Sigma}_{uu} = c\boldsymbol{\Sigma}_{xx}$ for constant c.

The above theorem shows that MEC is sufficient for asymptotic sign consistency, however it does not imply that the inverse is true. We can obtain a better idea about the necessary and sufficient conditions for sign consistency by considering the elastic net in the absence of model error i.e. $\boldsymbol{\epsilon} = \mathbf{0}$. In the finite sample measurement-error free situation with $\boldsymbol{\epsilon} = \mathbf{0}$, the elastic net/lasso

estimate is sign consistent if and only if the corresponding Irrepresentable Condition holds. We now consider the necessary and sufficient conditions for sign consistency in the noise free model with measurement-error.

Theorem 2. Consider the naive elastic net when there is no model error, $\boldsymbol{\epsilon} = \mathbf{0}$. Let the detectable set be defined as $S_0^{det} =$

$$\left\{ j : |\beta_j^0| > \frac{\lambda_1}{2} \left(\underset{||\tau_{S_0}||_\infty \leq 1}{sup} \left\| (\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I})^{-1} \left(\boldsymbol{\tau}_{S_0} + \frac{2\lambda_2}{\lambda_1} \boldsymbol{\beta}_{S_0}^0 \right) \right\|_\infty \right) + |v_j| \right\}$$

where $\boldsymbol{v} = (\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I})^{-1} \mathbf{C}_{wu}(S_0, S_0) \boldsymbol{\beta}_{S_0}^0$.

1. If the EIC-ME is satisfied and

$$\left(\mathbf{C}_{ww}(S_0^c, S_0)(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I})^{-1} \mathbf{C}_{wu}(S_0, S_0) - \mathbf{C}_{wu}(S_0^c, S_0) \right) \boldsymbol{\beta}_{S_0}^0 = 0,$$

   then $S_0^{det} \subseteq \hat{S}(\lambda_1, \lambda_2) \subseteq S_0$, where $\hat{S}(\lambda_1, \lambda_2)$ is the active set of the estimate under λ_1, λ_2.

2. Suppose $\hat{S}(\lambda_1, \lambda_2) = S_0 = S_0^{det}$. Then

$$\left\| \mathbf{C}_{ww}(S_0^c, S_0)(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I})^{-1} \left(\text{sign}(\boldsymbol{\beta}_{S_0}^0) + \frac{2\lambda_2}{\lambda_1} \boldsymbol{\beta}_{S_0}^0 \right) \right.$$
$$\left. + \frac{2}{\lambda_1} \left(\mathbf{C}_{ww}(S_0^c, S_0)(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I})^{-1} \mathbf{C}_{wu}(S_0, S_0) - \mathbf{C}_{wu}(S_0^c, S_0) \right) \boldsymbol{\beta}_{S_0}^0 \right\|_\infty \leq 1.$$

Proof. We consider the elastic net with $\boldsymbol{\epsilon} = \mathbf{0}$. Then $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^0$ and hence the elastic net is

$$\hat{\boldsymbol{\beta}}_{Naive-EN} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{n} ||\mathbf{W}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}^0||_2^2 + \lambda_2 ||\boldsymbol{\beta}||_2^2 + \lambda_1 ||\boldsymbol{\beta}||_1 \right\}.$$

Part 1. From the KKT conditions we have,

$$2 \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I} \right) (\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0) + 2\mathbf{C}_{ww}(S_0, S_0^c) \hat{\boldsymbol{\beta}}_{S_0^c}$$
$$+ 2 \left(\mathbf{C}_{wu}(S_0, S_0) + \lambda_2 \mathbf{I} \right) \boldsymbol{\beta}_{S_0}^0 = -\lambda_1 \hat{\boldsymbol{\tau}}_{S_0}$$

$$2\mathbf{C}_{ww}(S_0^c, S_0)(\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0) + 2\mathbf{C}_{ww}(S_0^c, S_0^c) \hat{\boldsymbol{\beta}}_{S_0^c} + 2\mathbf{C}_{wu}(S_0^c, S_0) \boldsymbol{\beta}_{S_0}^0 = -\lambda_1 \hat{\boldsymbol{\tau}}_{S_0^c}$$

where $||\hat{\boldsymbol{\tau}}||_\infty \leq 1$ and $\hat{\tau}_j = \text{sign}\left(\hat{\beta}_j\right)$ if $\beta_j^0 \neq 0$. Multiplying the first equation by $\hat{\boldsymbol{\beta}}_{S_0^c}^T \mathbf{C}_{ww}(S_0^c, S_0) \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1}$ and the second equation by $\hat{\boldsymbol{\beta}}_{S_0^c}^T$ gives,

$$\begin{aligned} 2\hat{\boldsymbol{\beta}}_{S_0^c}^T \mathbf{C}_{ww}(S_0^c, S_0)(\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0) + 2\hat{\boldsymbol{\beta}}_{S_0^c}^T \mathbf{C}_{ww}(S_0^c, S_0) \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \mathbf{C}_{ww}(S_0, S_0^c) \hat{\boldsymbol{\beta}}_{S_0^c} \\ + 2\hat{\boldsymbol{\beta}}_{S_0^c}^T \mathbf{C}_{ww}(S_0^c, S_0) \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \left(\mathbf{C}_{wu}(S_0, S_0) + \lambda_2 \mathbf{I}\right) \boldsymbol{\beta}_{S_0}^0 \\ = -\lambda_1 \hat{\boldsymbol{\beta}}_{S_0^c}^T \mathbf{C}_{ww}(S_0^c, S_0) \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \hat{\boldsymbol{\tau}}_{S_0}, \end{aligned}$$

$$2\hat{\boldsymbol{\beta}}_{S_0^c}^T \mathbf{C}_{ww}(S_0^c, S_0)(\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0) + 2\hat{\boldsymbol{\beta}}_{S_0^c}^T \mathbf{C}_{ww}(S_0^c, S_0^c) \hat{\boldsymbol{\beta}}_{S_0^c} + 2\hat{\boldsymbol{\beta}}_{S_0^c}^T \mathbf{C}_{wu}(S_0^c, S_0) \boldsymbol{\beta}_{S_0}^0 = -\lambda_1 \hat{\boldsymbol{\beta}}_{S_0^c}^T \hat{\boldsymbol{\tau}}_{S_0^c}.$$

Now subtracting the first of these equations from the second and re-arranging we get,

$$\begin{aligned} 2\hat{\boldsymbol{\beta}}_{S_0^c}^T \left(\mathbf{C}_{ww}(S_0^c, S_0^c) - \mathbf{C}_{ww}(S_0^c, S_0) \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \mathbf{C}_{ww}(S_0, S_0^c)\right) \hat{\boldsymbol{\beta}}_{S_0^c} \\ + 2\hat{\boldsymbol{\beta}}_{S_0^c}^T \left(\mathbf{C}_{wu}(S_0^c, S_0) - \mathbf{C}_{ww}(S_0^c, S_0) \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \mathbf{C}_{wu}(S_0, S_0)\right) \boldsymbol{\beta}_{S_0}^0 \\ = \lambda_1 \left(\hat{\boldsymbol{\beta}}_{S_0^c}^T \mathbf{C}_{ww}(S_0^c, S_0) \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \left(\hat{\boldsymbol{\tau}}_{S_0} + \frac{2\lambda_2}{\lambda_1} \boldsymbol{\beta}_{S_0}^0\right) - \hat{\boldsymbol{\beta}}_{S_0^c} \hat{\boldsymbol{\tau}}_{S_0^c}\right). \end{aligned}$$

The first term on the LHS is a positive semi-definite matrix. On the RHS,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{S_0^c}^T \mathbf{C}_{ww}(S_0^c, S_0) \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \left(\hat{\boldsymbol{\tau}}_{S_0} + \frac{2\lambda_2}{\lambda_1} \boldsymbol{\beta}_{S_0}^0\right) - ||\hat{\boldsymbol{\beta}}_{S_0^c}||_1 \leq \\ \left(\left\| \mathbf{C}_{ww}(S_0^c, S_0) \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \left(\hat{\boldsymbol{\tau}}_{S_0} + \frac{2\lambda_2}{\lambda_1} \boldsymbol{\beta}_{S_0}^0\right) \right\|_\infty - 1\right) ||\hat{\boldsymbol{\beta}}_{S_0^c}||_1 \leq 0 \end{aligned}$$

due to the EIC-ME condition. Finally,

$$\mathbf{C}_{wu}(S_0^c, S_0) - \mathbf{C}_{ww}(S_0^c, S_0) \left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \mathbf{C}_{wu}(S_0, S_0)$$

is assumed to be zero. Therefore, if $||\hat{\boldsymbol{\beta}}_{S_0^c}||_1 \neq 0$ then the LHS must be negative, which gives a contradiction. Hence, $\hat{\boldsymbol{\beta}}_{S_0^c} = 0$ which means the KKT conditions become,

$$2\left(\mathbf{C}_{ww}(S_0, S_0) + \lambda_2 \mathbf{I}\right)(\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0) + 2(\mathbf{C}_{wu}(S_0, S_0) + \lambda_2 \mathbf{I}) \boldsymbol{\beta}_{S_0}^0 = -\lambda_1 \hat{\boldsymbol{\tau}}_{S_0} \tag{6.6}$$

$$2\mathbf{C}_{ww}(S_0^c, S_0)(\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0) + 2\mathbf{C}_{wu}(S_0^c, S_0) \boldsymbol{\beta}_{S_0}^0 = -\lambda_1 \hat{\boldsymbol{\tau}}_{S_0^c}. \tag{6.7}$$

From the first of these we get,

$$\begin{aligned}|\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^0_{S_0}| = &\left|\frac{\lambda_1}{2}\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\left(\hat{\boldsymbol{\tau}}_{S_0}+\frac{2\lambda_2}{\lambda_1}\boldsymbol{\beta}^0_{S_0}\right)\right.\\ &\left.+\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\mathbf{C}_{wu}(S_0,S_0)\boldsymbol{\beta}^0_{S_0}\right|\\ &\leq \frac{\lambda_1}{2}\left(\underset{||\boldsymbol{\tau}||_\infty\leq 1}{sup}\left\|\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\left(\boldsymbol{\tau}_{S_0}+\frac{2\lambda_2}{\lambda_1}\boldsymbol{\beta}^0_{S_0}\right)\right\|_\infty\right)\mathbf{1}\\ &+|(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I})^{-1}\mathbf{C}_{wu}(S_0,S_0)\boldsymbol{\beta}^0_{S_0}|.\end{aligned}$$

Now if $j \in S_0^{det}$ and $\hat{\beta}_j = 0$, then we have

$$|\hat{\beta}_j - \beta^0_j| = |\beta^0_j| > \frac{\lambda_1}{2}\left(\underset{||\boldsymbol{\tau}||_\infty\leq 1}{sup}\left\|\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\left(\boldsymbol{\tau}_{S_0}+\frac{2\lambda_2}{\lambda_1}\boldsymbol{\beta}^0_{S_0}\right)\right\|_\infty\right)+|v_j|,$$

where $\boldsymbol{v} = (\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I})^{-1}\mathbf{C}_{wu}(S_0,S_0)\boldsymbol{\beta}^0_{S_0}$. This is a contradiction, so $\hat{\beta}_j \neq 0$ for $j \in S_0^{det}$.

Part 2. Now suppose $\hat{S} = S_0 = S_0^{det}$. Then we can assume $\text{sign}(\hat{\boldsymbol{\beta}}) = \text{sign}(\boldsymbol{\beta}^0)$. Now from the KKT conditions we have,

$$\begin{aligned}(\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^0_{S_0}) = &-\frac{\lambda_1}{2}\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\left(\hat{\boldsymbol{\tau}}_{S_0}+\frac{2\lambda_2}{\lambda_1}\boldsymbol{\beta}^0_{S_0}\right)\\ &-\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\mathbf{C}_{wu}(S_0,S_0)\boldsymbol{\beta}^0_{S_0}\end{aligned}$$

and substituting into (6.7) gives us,

$$\begin{aligned}&-\lambda_1\mathbf{C}_{ww}(S_0^c,S_0)\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\left(\hat{\boldsymbol{\tau}}_{S_0}+\frac{2\lambda_2}{\lambda_1}\boldsymbol{\beta}^0_{S_0}\right)\\ &-2\mathbf{C}_{ww}(S_0^c,S_0)\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\mathbf{C}_{wu}(S_0,S_0)\boldsymbol{\beta}^0_{S_0}+2\mathbf{C}_{wu}(S_0^c,S_0)\boldsymbol{\beta}^0_{S_0} = -\lambda_1\hat{\boldsymbol{\tau}}_{S_0^c}\\ &\implies\\ &\mathbf{C}_{ww}(S_0^c,S_0)\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\left(\hat{\boldsymbol{\tau}}_{S_0}+\frac{2\lambda_2}{\lambda_1}\boldsymbol{\beta}^0_{S_0}\right)\\ &+\frac{2}{\lambda_1}\left(\mathbf{C}_{ww}(S_0^c,S_0)\left(\mathbf{C}_{ww}(S_0,S_0)+\lambda_2\mathbf{I}\right)^{-1}\mathbf{C}_{wu}(S_0,S_0)-\mathbf{C}_{wu}(S_0^c,S_0)\right)\boldsymbol{\beta}^0_{S_0} = \hat{\boldsymbol{\tau}}_{S_0^c}.\end{aligned}$$

Bounding this gives the result in the second part of the theorem.

□

Hence the EIC-ME on its own is no longer a necessary condition for sign consistency in the presence of measurement-error. The vector $\boldsymbol{v}$ in Theorem 2 in-

volves the measurement-error and $\boldsymbol{\beta}^0$ and hence in the presence of measurement-error the elastic net will struggle to identify arbitrarily small coefficients. Therefore, in order for the naive elastic net to be sign consistent we also require the much stricter MEC condition to hold. We therefore need to adapt the elastic net to correct for the measurement-error.

6.3 THE CORRECTED ELASTIC NET

The corrected elastic net employs the same reasoning as the lasso with measurement-error correction and is based on the fact that the expectation of the loss function is biased. The objective function of the corrected elastic net is

$$\hat{\boldsymbol{\beta}}_{C-EN} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{n}||\mathbf{y} - \mathbf{W}\boldsymbol{\beta}||_2^2 - \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{uu}\boldsymbol{\beta} + \lambda_2||\boldsymbol{\beta}||_2^2 + \lambda_1||\boldsymbol{\beta}||_1 \right\}.$$

Note, if the covariance matrix of the measurement-error is diagonal, i.e. $\boldsymbol{\Sigma}_{uu} = \sigma_u^2\mathbf{I}$ then this reduces to

$$\hat{\boldsymbol{\beta}}_{C-EN} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{n}||\mathbf{y} - \mathbf{W}\boldsymbol{\beta}||_2^2 + (\lambda_2 - \sigma_u^2)\boldsymbol{\beta}^T\boldsymbol{\beta} + \lambda_1||\boldsymbol{\beta}||_1 \right\},$$

which, provided $\lambda_2 > \sigma_u^2$ is simply the elastic net with $(\lambda_2 - \sigma_u^2)$ replacing the usual ℓ_2 norm regularisation parameter. In cases where the loss function is non-convex (such as for the lasso when $\lambda_2 = 0$ and $p > n$) we need to restrict the parameter space of $\boldsymbol{\beta}$ to $\boldsymbol{\beta} : ||\boldsymbol{\beta}||_1 \leq R$ where R is some finite radius. This is done to avoid trivial solutions.

As with section 6.2.1 we now repeat the results from Sørensen et al. (2015) for the corrected lasso, with the relevant extensions to the elastic net. i.e. we now replace $(\mathbf{C}_{w,w} - \boldsymbol{\Sigma}_{u,u})$ and $(\mathbf{C}_{w,u} - \boldsymbol{\Sigma}_{u,u})$ with $(\mathbf{C}_{w,w} - \boldsymbol{\Sigma}_{u,u} + \lambda_2\mathbf{I})$ and $(\mathbf{C}_{w,u} - \boldsymbol{\Sigma}_{u,u} + \lambda_2\mathbf{I})$ respectively. We now require a new Irrepresentable condition for the measurement-error corrected elastic net.

Definition 6.3.1. Elastic Irrepresentable Condition for the corrected elastic net (EIC-CEN). There exists a constant $\theta \in [0, 1)$ such that,

$$\Big\| (\mathbf{C}_{ww}(S_0^c, S_0) - \mathbf{\Sigma}_{uu}(S_0^c, S_0)) \left(\mathbf{C}_{ww}(S_0, S_0) - \mathbf{\Sigma}_{uu}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \times \left(\text{sign}(\boldsymbol{\beta}^0_{S_0}) + 2\frac{\lambda_2}{\lambda_1}\boldsymbol{\beta}^0_{S_0}\right)\Big\|_\infty \leq \theta.$$

We can now assess the conditions for sign consistency in the measurement-error corrected elastic net.

Lemma 2. If we assume the EIC-CEN holds for constant θ, and, if required, the optimum estimate $\hat{\boldsymbol{\beta}}$ lies in the ℓ_1 ball with some finite radius R, then

$$\mathbb{P}\left(\text{sign}(\hat{\boldsymbol{\beta}}) = \text{sign}(\boldsymbol{\beta}^0)\right) \geq \mathbb{P}(C \cap D)$$

where the events C and D are defined as,

$$C = \Bigg\{ |\mathbf{Z}_5 \boldsymbol{\epsilon} - \mathbf{Z}_6 \boldsymbol{\beta}^0_{S_0}| < \sqrt{n}\left(|\boldsymbol{\beta}^0_{S_0}| - \frac{\lambda_1}{2}\left|\left(\mathbf{C}_{ww}(S_0, S_0) - \mathbf{\Sigma}_{uu}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \text{sign}(\boldsymbol{\beta}^0_{S_0})\right|\right)\Bigg\},$$

$$D = \left\{ |\mathbf{Z}_7 \boldsymbol{\epsilon} - \mathbf{Z}_8 \boldsymbol{\beta}^0_{S_0}| < \frac{\lambda_1}{2}(1-\theta)\mathbf{1} \right\},$$

where,

$$\mathbf{Z}_5 = \left(\mathbf{C}_{ww}(S_0, S_0) - \mathbf{\Sigma}_{uu}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \frac{\mathbf{W}^T_{S_0}}{\sqrt{n}}$$

$$\mathbf{Z}_6 = \sqrt{n}\left(\mathbf{C}_{ww}(S_0, S_0) - \mathbf{\Sigma}_{uu}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \left(\mathbf{C}_{wu}(S_0, S_0) - \mathbf{\Sigma}_{uu}(S_0, S_0) + \lambda_2 \mathbf{I}\right)$$

$$\mathbf{Z}_7 = \left(\mathbf{C}_{ww}(S_0^c, S_0) - \mathbf{\Sigma}_{uu}(S_0^c, S_0)\right)\left(\mathbf{C}_{ww}(S_0, S_0) - \mathbf{\Sigma}_{uu}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \frac{\mathbf{W}^T_{S_0}}{\sqrt{n}} - \frac{\mathbf{W}^T_{S_0^c}}{\sqrt{n}}$$

$$\mathbf{Z}_8 = \sqrt{n}\Big(\left(\mathbf{C}_{ww}(S_0^c, S_0) - \mathbf{\Sigma}_{uu}(S_0^c, S_0)\right)\left(\mathbf{C}_{ww}(S_0, S_0) - \mathbf{\Sigma}_{uu}(S_0, S_0) + \lambda_2 \mathbf{I}\right)^{-1} \left(\mathbf{C}_{wu}(S_0, S_0) - \mathbf{\Sigma}_{uu}(S_0, S_0)\right) - \left(\mathbf{C}_{wu}(S_0^c, S_0) - \mathbf{\Sigma}_{uu}(S_0^c, S_0)\right)\Big).$$

Proof. Again, as with Lemma 1, take $\hat{\gamma} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0$. The KKT conditions for the corrected elastic net are satisfied if there exists an optimal $\hat{\boldsymbol{\gamma}}$. If the solution lies in the feasible set i.e. $||\hat{\boldsymbol{\gamma}} + \boldsymbol{\beta}^0||_1 < R$, then

$$-\frac{2}{n}\boldsymbol{\epsilon}^T \mathbf{W} + 2(\mathbf{C}_{ww} - \mathbf{\Sigma}_{uu} + \lambda_2 \mathbf{I})\hat{\boldsymbol{\gamma}} + 2(\mathbf{C}_{wu} - \mathbf{\Sigma}_{uu} + \lambda_2 \mathbf{I})\boldsymbol{\beta}^0 + \lambda_1 \hat{\boldsymbol{\tau}} = 0,$$

where $\hat{\boldsymbol{\tau}}$ is defined as in the proof of Lemma 1. By the KKT conditions, the corrected elastic net will give sign consistent selection, i.e. $\text{sign}(\hat{\boldsymbol{\beta}}_{S_0}) = \text{sign}(\boldsymbol{\beta}^0_{S_0})$ and $\hat{\boldsymbol{\beta}}_{S_0^c} = \mathbf{0}$, if there exists a $\hat{\boldsymbol{\gamma}}$ and the following hold,

$$-\frac{\mathbf{W}^T_{S_0}}{\sqrt{n}}\boldsymbol{\epsilon} + \sqrt{n}(\mathbf{C}_{ww}(S_0, S_0) - \boldsymbol{\Sigma}_{uu}(S_0, S_0) + \lambda_2\mathbf{I})\hat{\boldsymbol{\gamma}}_{S_0} + \sqrt{n}(\mathbf{C}_{wu}(S_0, S_0) - \boldsymbol{\Sigma}_{uu}(S_0, S_0) + \lambda_2\mathbf{I})\boldsymbol{\beta}^0_{S_0} = -\frac{\lambda_1\sqrt{n}}{2}\text{sign}(\boldsymbol{\beta}^0_{S_0}) \tag{6.8}$$

$$|\hat{\boldsymbol{\gamma}}_{S_0}| < |\boldsymbol{\beta}^0_{S_0}| \tag{6.9}$$

$$\left| -\frac{\mathbf{W}^T_{S_0^c}}{\sqrt{n}}\boldsymbol{\epsilon} + \sqrt{n}\left(\mathbf{C}_{ww}(S_0^c, S_0) - \boldsymbol{\Sigma}_{uu}(S_0^c, S_0)\right)\hat{\boldsymbol{\gamma}}_{S_0} + \sqrt{n}\left(\mathbf{C}_{wu}(S_0^c, S_0) - \boldsymbol{\Sigma}_{uu}(S_0^c, S_0)\right)\boldsymbol{\beta}^0_{S_0} \right| \leq \frac{\lambda_1\sqrt{n}}{2}\mathbf{1}. \tag{6.10}$$

Again in the third condition the λ_2 terms drop from the inequality. The rest of the proof then follows the same reasoning as that of Lemma 1.

□

We now investigate the probability of asymptotic sign consistency for the corrected elastic net.

Theorem 3. Assume suitable regularity conditions, if λ_1,$\lambda_2 \to 0$ and $\lambda_1 n^{(1-c)/2} \to \infty$ as $n \to \infty$ for some $c \in [0, 1)$, then

$$\mathbb{P}\left(\text{sign}(\hat{\boldsymbol{\beta}}) = \text{sign}(\boldsymbol{\beta}^0)\right) = 1 - o(e^{-n^c}).$$

Proof. Again, assume the regularity conditions from Theorem 1.

$$1 - \mathbb{P}(C \cap D) \leq \mathbb{P}(C^c) + \mathbb{P}(D^c) \leq \sum_{j \in S_0} \mathbb{P}\left(|(\mathbf{Z}_5\boldsymbol{\epsilon})_j| \geq \sqrt{n}\left(\mathbf{c}_j - \frac{\lambda_1}{2}\mathbf{d}_j\right)\right) + \sum_{j \in S_0^c} \mathbb{P}\left(|(\mathbf{Z}_7\boldsymbol{\epsilon})_j - (\mathbf{Z}_8\boldsymbol{\beta}^0_{S_0})_j| \geq \frac{\lambda_1\sqrt{n}}{2}(1-\theta)\right)$$

where,

$$\begin{aligned}\mathbf{c} &= |\boldsymbol{\beta}^0_{S_0}| \\ &\quad - \left|(\mathbf{C}_{ww}(S_0,S_0) - \boldsymbol{\Sigma}_{uu}(S_0,S_0) + \lambda_2\mathbf{I})^{-1}\,(\mathbf{C}_{wu}(S_0,S_0) - \boldsymbol{\Sigma}_{uu}(S_0,S_0) + \lambda_2\mathbf{I})\boldsymbol{\beta}^0_{S_0}\right| \\ \mathbf{d} &= (\mathbf{C}_{ww}(S_0,S_0) - \boldsymbol{\Sigma}_{uu}(S_0,S_0) + \lambda_2\mathbf{I})^{-1}\left(\text{sign}(\boldsymbol{\beta}^0_{S_0})\right).\end{aligned}$$

Clearly, $\mathbf{Z}_5\boldsymbol{\epsilon}$ converges in distribution to a normal distribution with mean $\mathbf{0}$ and covariance $\sigma^2\boldsymbol{\Sigma}_{xx}(S_0,S_0)^{-1}\boldsymbol{\Sigma}_{ww}(S_0,S_0)\boldsymbol{\Sigma}_{xx}(S_0,S_0)^{-1}$, as $n \to \infty$. Therefore, for $j \in S_0$, there exists a finite κ such that $\mathbb{E}((\mathbf{Z}_5\boldsymbol{\epsilon})_j)^2 < \kappa^2$.

If our assumptions about λ_2 and the regularity conditions hold,

$$\mathbf{c} \to |\boldsymbol{\beta}^0_{S_0}|,$$

and

$$\mathbf{d} \to \boldsymbol{\Sigma}_{xx}(S_0,S_0)^{-1}\text{sign}(\boldsymbol{\beta}^0_{S_0}),\ \text{as } n \to \infty.$$

Also, $\mathbf{Z}_7\boldsymbol{\epsilon}$ converges in distribution to a normal distribution with mean $\mathbf{0}$ and covariance,

$$\sigma^2(\boldsymbol{\Sigma}_{xx}(S_0^c,S_0^c) - \boldsymbol{\Sigma}_{xx}(S_0^c,S_0)\boldsymbol{\Sigma}_{xx}(S_0,S_0)^{-1}\boldsymbol{\Sigma}_{xx}(S_0,S_0^c)).$$

The rest of the proof closely follows that of Theorem 1. The key exception uses the fact that $\sqrt{n}\mathbf{C}_{wu}$ is normally distributed with mean $\sqrt{n}\boldsymbol{\Sigma}_{wu} = \sqrt{n}\boldsymbol{\Sigma}_{uu}$ (Anderson, 2003) and $\mathbf{C}_{ww} - \boldsymbol{\Sigma}_{uu} \to \boldsymbol{\Sigma}_{xx}$ to show that, as $n \to \infty$, then

$$\begin{aligned}\mathbf{Z}_8 = \sqrt{n}\Big((\mathbf{C}_{ww}(S_0^c,S_0) - \boldsymbol{\Sigma}_{uu}(S_0^c,S_0))\,(\mathbf{C}_{ww}(S_0,S_0) - \boldsymbol{\Sigma}_{uu}(S_0,S_0) + \lambda_2\mathbf{I})^{-1} \\ (\mathbf{C}_{wu}(S_0,S_0) - \boldsymbol{\Sigma}_{uu}(S_0,S_0)) - (\mathbf{C}_{wu}(S_0^c,S_0) - \boldsymbol{\Sigma}_{uu}(S_0^c,S_0))\Big)\end{aligned} \tag{6.11}$$

is normally distributed with mean zero and finite variance. Therefore, $\mathbf{Z}_8\boldsymbol{\beta}^0_{S_0}$ is also normally distributed with mean zero and finite variance without the need for the MEC condition.

□

Therefore, assuming the EIC-CEN holds the corrected elastic net can give sign consistent selection without the need for the strict MEC condition. If the loss function of the corrected elastic net is non-convex, we require a strict condition of the parameter space $||\boldsymbol{\beta}||_1 < R$ since the boundary of the ℓ_1 ball with radius R may represent a local optimum if the objective function is decreasing as we cross the boundary (Sørensen et al., 2015). However, Loh and Wainwright (2012) demonstrated that the optimisation error for a local optimum is $O\left(s_0\sqrt{p/n}\right)$ and any local optima are contained in a small ℓ_1 ball around $\boldsymbol{\beta}^0$ when n is of sufficient size.

6.4 MULTIPLICATIVE MEASUREMENT-ERROR

We now consider the lasso and the elastic net in the presence of multiplicative measurement-error. Let $\mathbf{W} = \mathbf{X} \odot \mathbf{U}$ where $\mathbf{U}$ is the measurement-error, and $\odot$ denotes the element-wise multiplication operator. We can expand out the expected loss in a similar way to the additive measurement-error case.

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{n}||\mathbf{y} - \mathbf{W}\boldsymbol{\beta}||_2^2|\mathbf{X}, \mathbf{y}\right) &= \frac{1}{n}\mathbb{E}\left(||\mathbf{y} - (\mathbf{X} \odot \mathbf{U})\boldsymbol{\beta}||_2^2\right) \\
&= \frac{1}{n}\mathbb{E}\left(\mathbf{y}^T\mathbf{y} + \boldsymbol{\beta}^T(\mathbf{X} \odot \mathbf{U})^T(\mathbf{X} \odot \mathbf{U})\boldsymbol{\beta} - 2\boldsymbol{\beta}^T(\mathbf{X} \odot \mathbf{U})^T\mathbf{y}\right) \\
&= \frac{1}{n}\mathbf{y}^T\mathbf{y} + \frac{1}{n}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X} \odot \mathbb{E}\left(\mathbf{U}^T\mathbf{U}\right)\boldsymbol{\beta} - \frac{2}{n}\boldsymbol{\beta}^T(\mathbf{X} \odot \mathbb{E}(\mathbf{U}))^T\mathbf{y}.
\end{aligned}$$

The lasso and elastic net minimisations can be expressed as M-estimators in the form,

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p}\left\{\boldsymbol{\beta}^T\hat{\boldsymbol{\Gamma}}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\hat{\boldsymbol{\Lambda}} + \lambda_1||\boldsymbol{\beta}||_1\right\},$$

where $\hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\Lambda}}$ depend on the type of measurement-error and whether we implement a lasso or elastic net. For example, $\hat{\boldsymbol{\Gamma}}_{Lass} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ and $\hat{\boldsymbol{\Lambda}}_{Lass} = \frac{1}{n}\mathbf{X}^T\mathbf{y}$ for the

lasso with no measurement-error. In the additive measurement-error case the lasso takes

$$\hat{\Gamma}_{L-Ad} = \frac{1}{n}\mathbf{W}^T\mathbf{W} - \boldsymbol{\Sigma}_u, \;\; \hat{\Lambda}_{L-Ad} = \frac{1}{n}\mathbf{W}^T\mathbf{y},$$

and the elastic net has

$$\hat{\Gamma}_{EN-Ad} = \frac{1}{n}\mathbf{W}^T\mathbf{W} - \boldsymbol{\Sigma}_u + \lambda_2\mathbf{I}, \;\; \hat{\Lambda}_{EN-Ad} = \frac{1}{n}\mathbf{W}^T\mathbf{y}.$$

Using the expansion of the loss function it is simple to see that a correction for the multiplicative measurement-error in the lasso can be obtained by setting,

$$\hat{\Gamma}_{L-Mult} = \frac{1}{n}\mathbf{W}^T\mathbf{W} \oslash \mathbb{E}\left(\mathbf{U}^T\mathbf{U}\right), \;\; \hat{\Lambda}_{L-Mult} = \frac{1}{n}\mathbf{W}^T\mathbf{y} \oslash \mathbb{E}(\mathbf{U}),$$

where $\oslash$ denotes the element-wise division. If we instead want the corresponding elastic net formulation then we set

$$\hat{\Gamma}_{EN-Mult} = \frac{1}{n}\mathbf{W}^T\mathbf{W} \oslash \mathbb{E}\left(\mathbf{U}^T\mathbf{U}\right) + \lambda_2\mathbf{I},$$

with $\hat{\Lambda}_{EN-Mult}$ remaining the same as $\hat{\Lambda}_{L-Mult}$. It is easy to see that $\hat{\Gamma}$ and $\hat{\Lambda}$ are set to be unbiased estimators of $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta}$ respectively.

6.4.1 *Log-normal measurement-error*

For multiplicative measurement-error we will commonly assume that the error follows a log-normal distribution (see section 5.2). For log-normal measurement-error, $\mathbf{U} \sim \text{lnN}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_u)$, it seems reasonable that we will often assume $\boldsymbol{\mu}_p = \mathbf{0}_p$ and $\boldsymbol{\Sigma}_u = \sigma_u^2\mathbf{I}_p$ for small σ_u^2 (i.e. $0 < \sigma_u^2 < 1$). Therefore assuming each row, k, of the multiplicative measurement-error matrix $\mathbf{U}$ is drawn from distribution $\mathbf{u}_k \sim \text{lnN}(\mathbf{0}_p, \sigma_u^2\mathbf{I}_p)$, the expectation of $\mathbf{U}$ will be

$$\mathbb{E}(\mathbf{U})_i = e^{\frac{1}{2}\sigma_u^2}$$

which is close to 1 when σ_u^2 is small. As well as the $\mathbb{E}(\mathbf{U})$ we also require the value of $\mathbb{E}\left(\mathbf{U}^T\mathbf{U}\right)$ for the correction method. Rearranging the variance formula gives,

$$\begin{aligned}\mathbb{E}\left(\mathbf{U}^T\mathbf{U}\right)_{ij} &= \mathrm{Var}(\mathbf{U})_{ij} + (\mathbb{E}(\mathbf{U}))_{ij}^2 \\ &= e^{\frac{1}{2}(\mathbf{\Sigma}_{u,ii}+\mathbf{\Sigma}_{u,jj})}\left(e^{\mathbf{\Sigma}_{u,ij}}-1\right) + e^{\frac{1}{2}(\mathbf{\Sigma}_{u,ii}+\mathbf{\Sigma}_{u,jj})} \\ &= e^{\sigma_u^2}\left(e^{\mathbf{\Sigma}_{u,ij}}-1\right) + e^{\sigma_u^2} = e^{\sigma_u^2+\mathbf{\Sigma}_{u,ij}}.\end{aligned}$$

Hence, on the lead diagonal, when $i = j$, $\mathbb{E}\left(\mathbf{U}^T\mathbf{U}\right)_{ii} = e^{2\sigma_u^2}$ and at the off-diagonals, $\mathbb{E}\left(\mathbf{U}^T\mathbf{U}\right)_{ij} = e^{\sigma_u^2}$.

For example assume each row $\mathbf{u}_k$, $k = 1, ..., n$ is distributed $\mathbf{u}_k = (u_{k,1}, u_{k,2}, ..., u_{k,p}) \sim \exp\left(N(\mathbf{0}_p, \mathbf{\Sigma}_u)\right)$, where $\mathbf{\Sigma}_u$ is a diagonal $p \times p$ matrix with the lead diagonal taking values $(\sigma_{u,1}^2, \sigma_{u,2}^2, ..., \sigma_{u,p}^2)$. If $\mathbf{W} = \mathbf{X} \odot \mathbf{U}$ is the measurement-error contaminated matrix then,

$$\mathbf{W} = \begin{bmatrix} x_{1,1}u_{1,1} & x_{1,2}u_{1,2} & \cdots & x_{1,p}u_{1,p} \\ x_{2,1}u_{2,1} & x_{2,2}u_{2,2} & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ x_{n,1}u_{n,1} & \cdots & \cdots & x_{n,p}u_{n,p} \end{bmatrix}.$$

Taking the transpose and multiplying by itself, we then have

$$\mathbf{W}^T\mathbf{W} = \begin{bmatrix} \sum_{k=1}^n x_{k,1}^2u_{k,1}^2 & \sum_{k=1}^n x_{k,1}x_{k,2}u_{k,1}u_{k,2} & \cdots & \sum_{k=1}^n x_{k,1}x_{k,p}u_{k,1}u_{k,p} \\ \sum_{k=1}^n x_{k,2}x_{k,1}u_{k,2}u_{k,1} & \sum_{k=1}^n x_{k,2}^2u_{k,2}^2 & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ \sum_{k=1}^n x_{k,p}x_{k,1}u_{k,p}u_{k,1} & \cdots & \cdots & \sum_{k=1}^n x_{k,p}^2u_{k,p}^2 \end{bmatrix}.$$

Now taking expectations, using the fact that

$\mathbb{E}(u_{1,i}) = \mathbb{E}(u_{2,i}) = \cdots = \mathbb{E}(u_{n,i})$,

$$\mathbb{E}\left(\mathbf{W}^T\mathbf{W}\right) =$$
$$\begin{bmatrix} \left(\sum_{k=1}^n x_{k,1}^2\right)\mathbb{E}\left(u_1^2\right) & \left(\sum_{k=1}^n x_{k,1}x_{k,2}\right)\mathbb{E}(u_1u_2) & \cdots & \left(\sum_{k=1}^n x_{k,1}x_{k,p}\right)\mathbb{E}(u_1u_p) \\ \left(\sum_{k=1}^n x_{k,2}x_{k,1}\right)\mathbb{E}(u_2u_1) & \left(\sum_{k=1}^n x_{k,2}^2\right)\mathbb{E}\left(u_2\right)^2 & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ \left(\sum_{k=1}^n x_{k,p}x_{k,1}\right)\mathbb{E}(u_pu_1) & \cdots & \cdots & \left(\sum_{k=1}^n x_{k,p}^2\right)\mathbb{E}\left(u_p^2\right) \end{bmatrix}.$$

Using the expectation and variance formulae for the log-normal distribution above,

$$\begin{aligned} \mathbb{E}\left(u_i^2\right) &= \mathrm{Var}(u_i) + \mathbb{E}(u_i)^2 \\ &= e^{\sigma_{u,i}^2}\left(e^{\sigma_{u,i}^2} - 1\right) + \left(e^{\frac{1}{2}\sigma_{u,i}^2}\right)^2 \\ &= e^{2\sigma_{u,i}^2} \end{aligned}$$

$$\begin{aligned} \mathbb{E}(u_iu_j) &= \mathrm{Cov}(u_i, u_j) + \mathbb{E}(u_i)\mathbb{E}(u_j) \\ &= 0 + e^{\frac{1}{2}\sigma_{u,i}^2}e^{\frac{1}{2}\sigma_{u,j}^2} = e^{\frac{1}{2}(\sigma_{u,i}^2+\sigma_{u,j}^2)}, \end{aligned}$$

and hence

$$\mathbb{E}\left(\mathbf{W}^T\mathbf{W}\right) = \left(\mathbf{X}^T\mathbf{X}\right) \odot \mathbf{E}_{uu},$$

where

$$\mathbf{E}_{uu} = \begin{bmatrix} e^{2\sigma_{u,1}^2} & e^{\frac{1}{2}(\sigma_{u,1}^2+\sigma_{u,2}^2)} & \dots & e^{\frac{1}{2}(\sigma_{u,1}^2+\sigma_{u,p}^2)} \\ e^{\frac{1}{2}(\sigma_{u,1}^2+\sigma_{u,2}^2)} & e^{2\sigma_{u,2}^2} & \dots & \vdots \\ \vdots & \dots & \ddots & \vdots \\ e^{\frac{1}{2}(\sigma_{u,1}^2+\sigma_{u,p}^2)} & \dots & \dots & e^{2\sigma_{u,p}^2} \end{bmatrix}.$$

Similarly,

$$\mathbf{W}^T\mathbf{y} = \begin{bmatrix} \sum_{k=1}^n x_{k,1}u_{k_1}y_k \\ \sum_{k=1}^n x_{k,2}u_{k_2}y_k \\ \vdots \\ \sum_{k=1}^n x_{k,p}u_{k_p}y_k \end{bmatrix},$$

and

$$\mathbb{E}\left(\mathbf{W}^T\mathbf{y}\right) = \begin{bmatrix} \left(\sum_{k=1}^n x_{k,1}y_k\right)\mathbb{E}(u_1) \\ \left(\sum_{k=1}^n x_{k,2}y_k\right)\mathbb{E}(u_2) \\ \vdots \\ \left(\sum_{k=1}^n x_{k,p}y_k\right)\mathbb{E}(u_p) \end{bmatrix}$$
$$= \left(\mathbf{X}^T\mathbf{y}\right) \odot \mathbf{E}_u$$

where $\mathbf{E}_u = \left(e^{\frac{1}{2}\sigma^2_{u,1}}, e^{\frac{1}{2}\sigma^2_{u,2}}, \cdots, e^{\frac{1}{2}\sigma^2_{u,p}}\right)^T$ is simply the column vector of the expected values of $u_i,\ i = 1, \cdots, p$.

6.5 SIMULATIONS

6.5.1 *Additive measurement-error*

The corrected lasso was tested on the measurement-error Monte Carlo dataset used in section 3.3. The objective function was optimised for a range of penalty levels and small covariates were set to zero using a threshold previously chosen by BIC. The best value of lambda was then chosen by BIC, before the model was refitted using only the non-zero covariates.

Tables 12-13 record the mean and standard deviation for the bias, true and false positives, and the proportion of true signal detected. The corrected lasso/elastic net estimates show a considerable drop in false positives compared to the naive estimates but also a small reduction in average true positives. For the lower measurement-error levels the bias is lower and across all measurement-error levels, the strength at the true locations is stronger. The larger bias for higher measurement-error is due to the false positives becoming stronger along with the true positives.

σ_u^2	Method	Bias	True +	False +	prop true
0	Lasso	0.72 (0.26)	4.78 (0.46)	22.88 (13.08)	0.95 (0.06)
0.1	Naive	3.44 (1.70)	4.29 (0.76)	24.55 (17.41)	0.75 (0.07)
	CorLass	1.44 (0.76)	3.96 (0.84)	1.54 (1.81)	0.89 (0.08)
0.2	Naive	4.71 (2.31)	4.07 (0.81)	23.16 (16.67)	0.63 (0.07)
	CorLass	3.68 (1.77)	3.64 (0.90)	5.53 (2.93)	0.78 (0.11)
0.4	Naive	6.21 (3.08)	3.79 (0.89)	22.72 (19.64)	0.47 (0.07)
	CorLass	6.55 (3.01)	3.47 (0.86)	8.84 (3.08)	0.64 (0.10)
0.6	Naive	6.65 (2.93)	3.50 (0.93)	18.26 (14.82)	0.37 (0.07)
	CorLass	9.35 (4.16)	3.18 (0.89)	12.75 (3.64)	0.53 (0.10)

Table 12: Results for naive and corrected lasso estimates for different measurement-error levels (standard deviation in brackets). Bias, true positives, false positives and proportion of signal detected at true locations. The number of true positives in the model is 5.

σ_u^2	Method	Bias	True +	False +	prop true
0	E-net	1.10 (0.34)	4.78 (0.48)	36.38 (13.33)	0.94 (0.07)
0.1	Naive	4.61 (2.17)	4.30 (0.77)	37.26 (19.62)	0.70 (0.07)
	CorEN	1.85 (0.94)	3.96 (0.84)	2.60 (2.31)	0.87 (0.08)
0.2	Naive	5.84 (2.78)	4.11 (0.79)	34.30 (19.98)	0.58 (0.07)
	CorEN	3.76 (2.04)	3.67 (0.88)	5.70 (4.11)	0.78 (0.10)
0.4	Naive	7.26 (3.65)	3.86 (0.86)	31.63 (22.12)	0.43 (0.07)
	CorEN	7.18 (3.17)	3.49 (0.90)	10.64 (4.35)	0.60 (0.11)
0.6	Naive	7.79 (3.62)	3.60 (0.92)	28.30 (21.33)	0.33 (0.07)
	CorEN	10.00 (4.17)	3.16 (0.93)	14.75 (4.41)	0.47 (0.10)

Table 13: Results for naive and corrected elastic net estimates for different measurement-error levels (standard deviation in brackets). Bias, true positives, false positives and proportion of signal detected at true locations. The number of true positives in the model is 5.

6.5.2 *Multiplicative measurement-error*

Moving on to multiplicative noise for each of the four levels of error variance ($\sigma_u^2 \in \{0.01, 0.04, 0.16, 0.25\}$), the measurement-error was simulated from a log-normal distribution with mean 1. i.e. $\mathbf{U} \sim \text{lnN}\left(-\frac{\sigma_u^2}{2}\mathbf{1}_p, \sigma_u^2\mathbf{I}_p\right)$. The corrected estimate was fitted for both the elastic net and lasso cases, with the naive estimate used as a starting point for the optimisation. The procedure was the same as the additive case in that the objective function was optimised twice, the second time being with just the non-zero covariates. Table 14 records the

bias, true positives, false positives and proportion of signal detected at the true positives for both the naive and corrected estimates. For each, the standard deviation is in parenthesis.

Method	σ_u^2	0.01	0.04	0.16	0.25
Lass (Naive)	Bias	1.33 (0.53)	2.32 (1.22)	4.47 (2.13)	5.41 (2.68)
	TP	4.76 (0.51)	4.57 (0.66)	4.20 (0.82)	4.12 (0.82)
	FP	25.59 (15.37)	24.56 (16.23)	23.87 (17.63)	24.30 (19.69)
	Prop true	0.91 (0.03)	0.83 (0.05)	0.65 (0.07)	0.57 (0.08)
Lass (Cor)	Bias	0.49 (0.23)	1.46 (1.11)	5.41 (3.36)	6.81 (3.27)
	TP	4.67 (0.58)	4.44 (0.71)	4.12 (0.84)	4.01 (0.85)
	FP	1.85 (2.38)	4.14 (4.04)	14.34 (8.76)	17.80 (8.64)
	Prop true	0.99 (0.02)	0.95 (0.06)	0.95 (0.24)	0.85 (0.21)
EN (Naive)	Bias	1.96 (0.69)	3.33 (1.57)	5.72 (2.62)	6.52 (3.10)
	TP	4.75 (0.51)	4.58 (0.64)	4.26 (0.80)	4.13 (0.85)
	FP	38.34 (14.73)	39.15 (18.60)	35.77 (19.79)	33.50 (21.55)
	Prop true	0.88 (0.03)	0.80 (0.05)	0.61 (0.07)	0.52 (0.08)
EN (Cor)	Bias	0.57 (0.29)	2.25 (1.19)	6.33 (3.17)	8.57 (4.39)
	TP	4.64 (0.60)	4.50 (0.68)	4.12 (0.83)	4.04 (0.85)
	FP	2.40 (3.92)	9.49 (4.28)	19.63 (8.54)	25.02 (9.40)
	Prop true	0.98 (0.02)	0.94 (0.05)	0.91 (0.20)	0.86 (0.25)

Table 14: Results for naive and corrected lasso/elastic net estimates for different multiplicative measurement-error levels. Bias, true positives, false positives and proportion of signal detected at true locations.

From Table 14 there are a number of trends that can be identified. Firstly, the bias and number of false positives both increase significantly for the higher measurement-error levels. On the other hand, the mean number of true positives shows a slight decrease as we increase the measurement-error. The affect on the selection of true positives is one factor in the higher bias of the larger

measurement-error cases. The increase in both the number and strength of the false positives will also contribute to the deviation from the true model.

Possibly the most interesting results in the table correspond to the performance of the corrected estimate at the true, non-zero, parameters as seen in the "prop true" rows.

For the lower levels of measurement-error the corrected estimates show consistently good results as the method has reliably managed to reconstruct the majority of the true signal. For the higher levels of measurement-error, the mean remains high, but the variability shows a significant increase. Closer inspection of this increase in variability shows that in some cases we get significant overcorrection, where one or more of the true parameters will be assigned a value that is up to twice its actual value. This will of course also contribute to the bias.

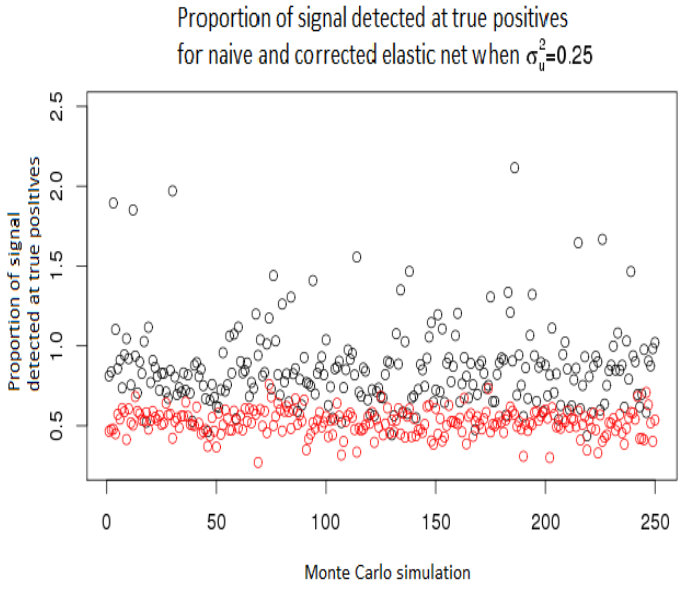

Figure 33: Proportion of signal detected at true positives in the naive (red) and corrected (black) estimates for the elastic net when $\sigma_u^2 = 0.25$ for 250 Monte Carlo simulations.

This overcorrection is actually linked to the level of penalisation included in the corrected model. Investigation into the effect of the penalty level in the scaling of the parameters shows that for the larger measurement-error levels, more penalisation was required to get the corrected estimates to closer match the true parameters, whereas for the lower measurement-error levels a small amount of penalisation usually gives good results. Despite the presence of overcorrection in some cases, instances of significant overestimation are usually outliers (see Fig. 33).

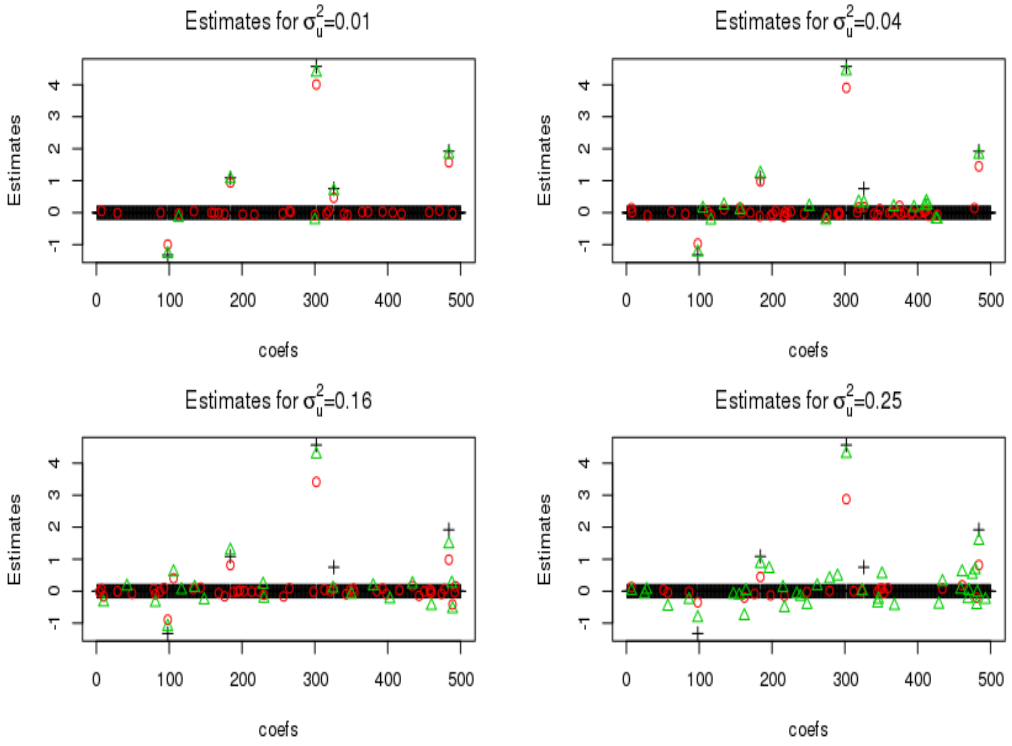

Figure 34: Naive (red) and corrected (green) estimates for the elastic net when, $\sigma_u^2 = 0.01, 0.04, 0.16$ and 0.25. True values in black. Note: only non-zero parameter estimates are plotted for the naive and corrected estimates.

Indeed the performance of the corrected lasso and elastic net in this simulation is usually a significant improvement over its naive counterpart. One such example of the naive and corrected estimates for the elastic net is shown in Fig.

34. For the lower measurement-error levels the corrected elastic net almost perfectly corrects its naive counterpart (table 14). For the more noisy cases, we see more influential false positives, however at the true positives, the effect of the correction is particularly evident.

In the simulations in this chapter we have seen that the corrected elastic net is an improvement on the naive estimate when the parameters are independent. The selection of the corrected elastic net under these circumstances appears to be good, with the exception that small covariates can sometimes be excluded. The case of the highly correlated MEG covariates will provide a greater challenge however. The corrected elastic net will be applied to real MEG data in chapter 8.

Our conclusions from the Monte Carlo studies must be tempered with the knowledge that, due to the relatively small number of samples included, the standard deviations on a number of the results are very high. Despite the appearance that the corrected estimates represent an improvement over the naive, the standard error levels are such that we cannot be certain that the difference is statistically significant. Therefore, in order to verify that a difference between the naive and corrected estimates exits, we would have to increase the sample size (and hence the computation time) substantially.

7

MEASUREMENT-ERROR SIMULATION RESULTS

In chapters 3-6 we have outlined a number of methods to correct for measurement-error in sparse regression methods. In order to test these methods against the naive estimate for a variety of measurement-error levels, Monte Carlo simulations were performed consisting of 500 parameters with 5 non-zero parameters. In order to compare how the measurement-error methods perform in relation to each other we now collect the results from the simulations according to the relevant sparse method. For each sparse method, we include the results from 250 Monte Carlo simulations with additive measurement-error for the naive estimate along with the SIMEX estimate, conditional score method, and where appropriate (for the lasso and elastic net) the corrected elastic net.

For comparison, the sparse estimate in the measurement-error free case will be included in order to demonstrate the performance under ideal conditions. In addition, for ease of reference, the best performing measurement-error method in terms of bias, true positives, false positives and the proportion of true activity detected will be indicated by bold font.

7.1 ADDITIVE

Table 15 contains the information for all the measurement-error corrections for the lasso. The best performing method for each level of measurement-error and category of performance is identified by emboldened font. From this we can identify recurring pattens in the performance of each method. It is also evident that there is a divide between the lower and higher levels of measurement-error.

σ_u^2	Method	Bias	True +	False +	prop true
0	Lasso	0.72 (0.26)	4.78 (0.46)	22.88 (13.08)	0.95 (0.06)
0.1	Naive	3.44 (1.70)	4.29 (0.76)	24.55 (17.41)	0.75 (0.07)
	SIMEX	4.89 (3.34)	**4.38 (0.74)**	50.97 (24.43)	**0.90 (0.10)**
	CorLass	**1.44 (0.76)**	3.96 (0.84)	**1.54 (1.81)**	0.89 (0.08)
	Sparse CS	4.27 (2.34)	4.29 (0.76)	24.55 (17.41)	0.86 (0.08)
0.2	Naive	4.71 (2.31)	4.07 (0.81)	23.16 (16.67)	0.63 (0.07)
	SIMEX	6.45 (4.38)	**4.22 (0.77)**	66.78 (26.64)	**0.83 (0.10)**
	CorLass	**3.68 (1.77)**	3.64 (0.90)	**5.53 (2.93)**	0.78 (0.11)
	Sparse CS	5.44 (3.02)	4.07 (0.81)	23.16 (16.67)	0.80 (0.10)
0.4	Naive	**6.21 (3.08)**	3.79 (0.89)	22.72 (19.64)	0.47 (0.07)
	SIMEX	8.40 (6.25)	**4.07 (0.84)**	69.87 (27.81)	**0.72 (0.13)**
	CorLass	6.55 (3.01)	3.47 (0.86)	**8.84 (3.08)**	0.64 (0.10)
	Sparse CS	6.59 (3.81)	3.80 (0.89)	22.80 (19.65)	**0.72 (0.11)**
0.6	Naive	**6.65 (2.93)**	3.50 (0.93)	18.26 (14.82)	0.37 (0.07)
	SIMEX	8.46 (5.24)	**3.86 (0.88)**	69.09 (29.80)	0.61 (0.13)
	CorLass	9.35 (4.16)	3.18 (0.89)	**12.75 (3.64)**	0.53 (0.10)
	Sparse CS	6.75 (3.61)	3.50 (0.93)	18.22 (14.82)	**0.67 (0.12)**

Table 15: Results summary for Lasso on 250 Monte Carlo simulations.

For low measurement-error the corrected lasso gives the lowest bias, however when the measurement-error increases the naive estimate is the least biased. This is because the corrected estimates tend to give greater signal weight to the false positives as well as the true positives. There is very little to separate the various estimates in the true positives as the means all fall within one standard deviation of the others. The corrected lasso easily performs the best in terms of false positives, whereas SIMEX introduces a large number of false positives to the model. However SIMEX does perform the best in correcting the estimates at the true covariates. Plotting the bias and true proportion against the measurement-error variance for each of the methods in the figures below we can obtain some other interesting conclusions. Firstly, from the bias plot (Fig. 35) it appears that the corrected lasso has an almost linear relationship with the bias as it deteriorates much more dramatically that the other estimates, where the bias curves begin to flatten out much sooner. Yet, for low measurement-error it is the best performing. This suggests that when we increase the error the estimate lends greater weight to the (increasing) false positives.

From the second figure, the corrected estimates all produce a significant improvement to the strength at the true locations. For the lower values of the measurement-error especially, there is little to separate the corrected estimates at the true covariates. As the lasso is a special case of the elastic net, we unsurprisingly see the same patterns for the elastic net estimates in Table 16 and Figs. 37-38. The performance of the elastic net is remarkably similar, yet not quite equal to the lasso. Across the measurement-error levels, the lasso has lower bias and fewer false positives. Therefore the elastic net's grouping property adds very little for data with uncorrelated covariates. In fact, for independent covariates the ℓ_2 norm term results in poorer selection performance.

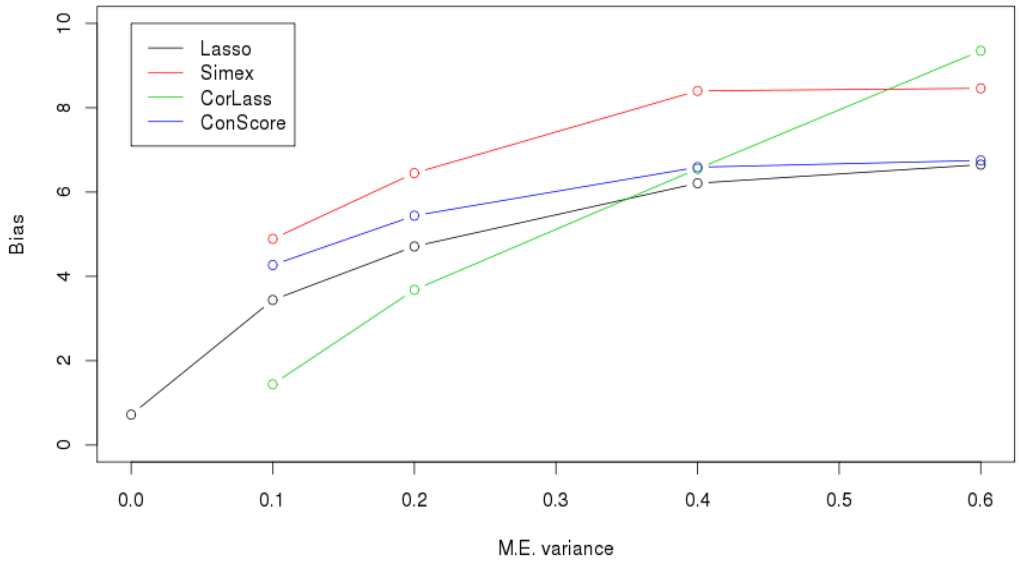

Figure 35: Bias of the lasso estimates and its corrections.

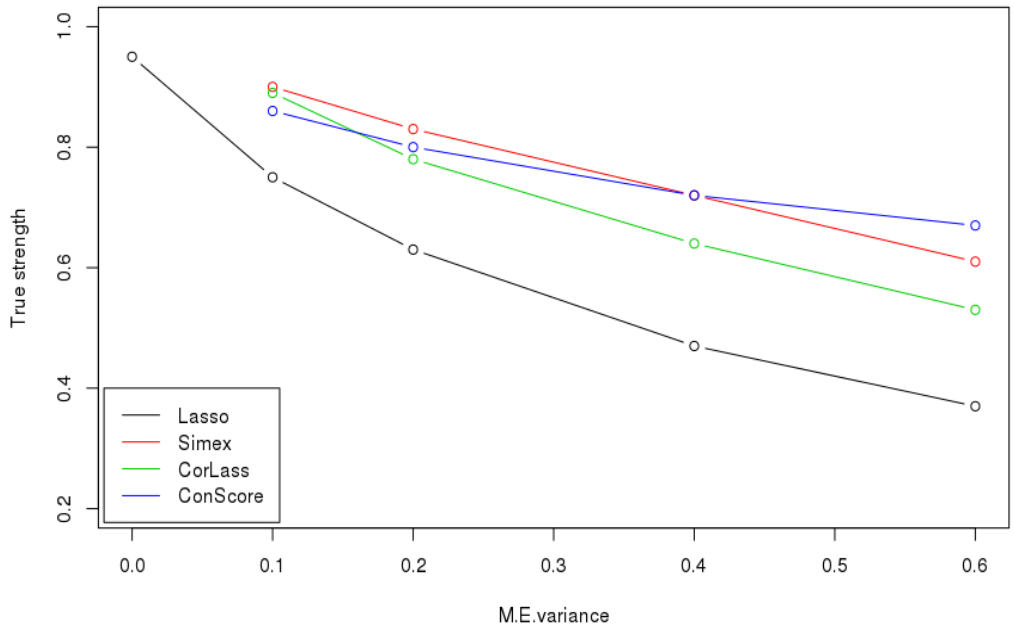

Figure 36: Strength at true locations for the lasso estimates and its corrections.

σ_u^2	Method	Bias	True +	False +	prop true
0	E-net	1.10 (0.34)	4.78 (0.48)	36.38 (13.33)	0.94 (0.07)
0.1	Naive	4.61 (2.17)	4.30 (0.77)	37.26 (19.62)	0.70 (0.07)
	SIMEX	6.76 (4.21)	**4.40 (0.71)**	87.41 (28.55)	0.86 (0.10)
	CorEN	**1.85 (0.94)**	3.96 (0.84)	**2.60 (2.31)**	**0.87 (0.08)**
	Sparse CS	5.70 (2.56)	4.30 (0.76)	37.36 (19.60)	0.82 (0.08)
0.2	Naive	5.84 (2.78)	4.11 (0.79)	34.30 (19.98)	0.58 (0.07)
	SIMEX	8.35 (5.18)	**4.31 (0.76)**	92.91 (30.33)	**0.78 (0.10)**
	CorEN	**3.76 (2.04)**	3.67 (0.88)	**5.70 (4.11)**	**0.78 (0.10)**
	Sparse CS	6.88 (3.28)	4.11 (0.79)	34.35 (20.00)	0.75 (0.10)
0.4	Naive	7.26 (3.65)	3.86 (0.86)	31.63 (22.12)	0.43 (0.07)
	SIMEX	10.47 (7.58)	**4.10 (0.79)**	94.88 (32.62)	0.66 (0.14)
	CorEN	**7.18 (3.17)**	3.49 (0.90)	**10.64 (4.35)**	0.60 (0.11)
	Sparse CS	8.18 (5.51)	3.86 (0.86)	31.71 (22.13)	**0.67 (0.12)**
0.6	Naive	**7.79 (3.62)**	3.60 (0.92)	28.30 (21.33)	0.33 (0.07)
	SIMEX	10.95 (7.00)	**3.89 (0.89)**	94.23 (35.74)	0.55 (0.13)
	CorEN	10.00 (4.17)	3.16 (0.93)	**14.75 (4.41)**	0.47 (0.10)
	Sparse CS	8.50 (4.52)	3.59 (0.92)	28.35 (21.39)	**0.62 (0.14)**

Table 16: Results summary for elastic net on 250 Monte Carlo simulations.

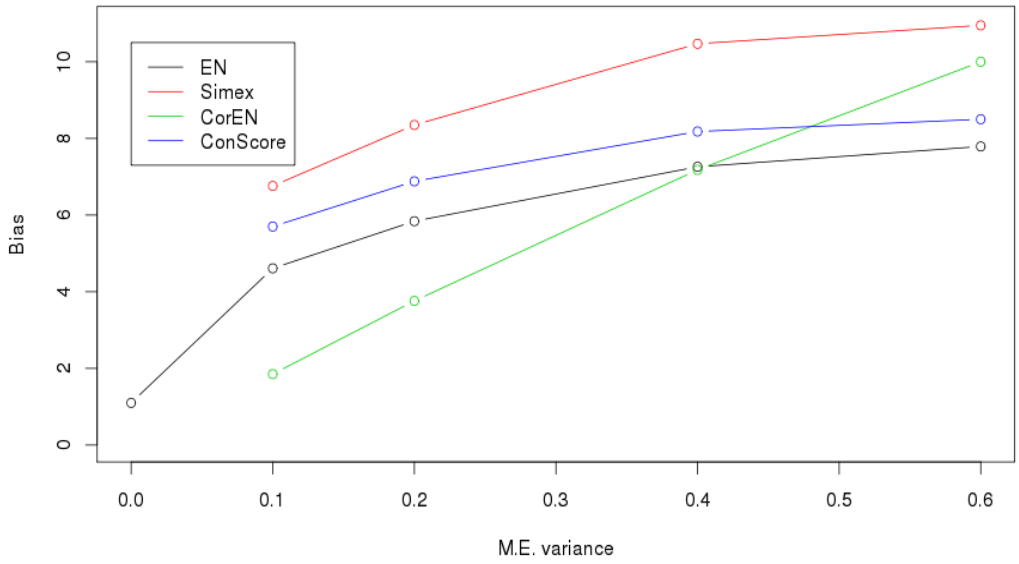

Figure 37: Bias of the elastic net estimates and its corrections.

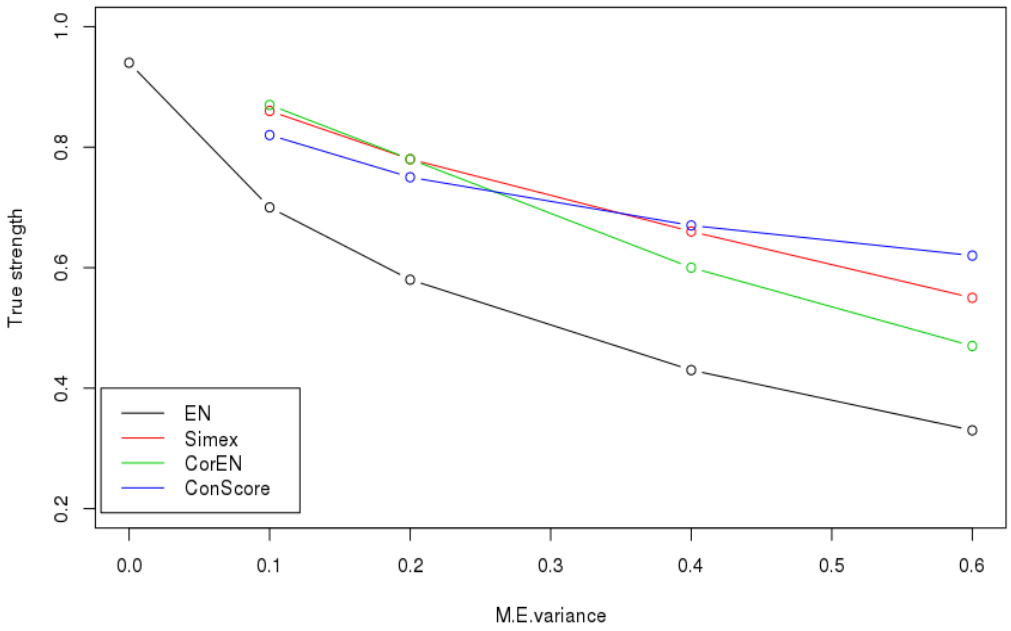

Figure 38: Strength at true locations for the elastic net estimates and its corrections.

σ_u^2	Method	Bias	True +	False +	prop true
0	SRL	4.44 (2.70)	3.32 (0.80)	0 (0.06)	0.45 (0.22)
0.1	Naive	5.49 (2.54)	2.70 (0.69)	**0 (0)**	0.30 (0.14)
	SIMEX	4.81 (2.59)	**2.94 (0.71)**	0.11 (0.32)	0.40 (0.19)
	Sparse CS	**3.22 (2.67)**	2.70 (0.69)	0.04 (0.28)	**0.81 (0.19)**
0.2	Naive	6.03 (2.59)	2.31 (0.68)	**0 (0.06)**	0.23 (0.12)
	SIMEX	5.21 (2.57)	**2.64 (0.72)**	0.20 (0.45)	0.34 (0.16)
	Sparse CS	**4.00 (3.00)**	2.32 (0.68)	0.20 (0.60)	**0.75 (0.24)**
0.4	Naive	6.62 (2.64)	1.86 (0.63)	**0.01 (0.09)**	0.15 (0.08)
	SIMEX	5.81 (2.60)	**2.32 (0.74)**	0.33 (0.56)	0.26 (0.14)
	Sparse CS	**5.18 (3.36)**	1.86 (0.63)	0.54 (0.88)	**0.68 (0.28)**
0.6	Naive	6.96 (2.71)	1.57 (0.67)	**0 (0.06)**	0.10 (0.07)
	SIMEX	6.25 (2.70)	**1.98 (0.72)**	0.33 (0.61)	0.20 (0.12)
	Sparse CS	**5.99 (4.33)**	1.58 (0.67)	0.87 (0.99)	**0.60 (0.31)**

Table 17: Results summary for Square root lasso on 250 Monte Carlo simulations.

The SIMEX estimate for the SRL gives a marginally higher average number of true positives, but it is not significantly better than the other methods and still averages less than 3 (out of 5) for even the smallest level of measurement-error. The sparse conditional score gives the best results in terms of bias and strength of the true covariates. The increase in the true strength (i.e. the proportion of signal at true, non-zero values, $\frac{||\hat{\boldsymbol{\beta}}_{S_0}||_1}{||\boldsymbol{\beta}^0_{S_0}||_1}$) suggests that even though the true positives are low, the strongest covariates have been correctly selected.

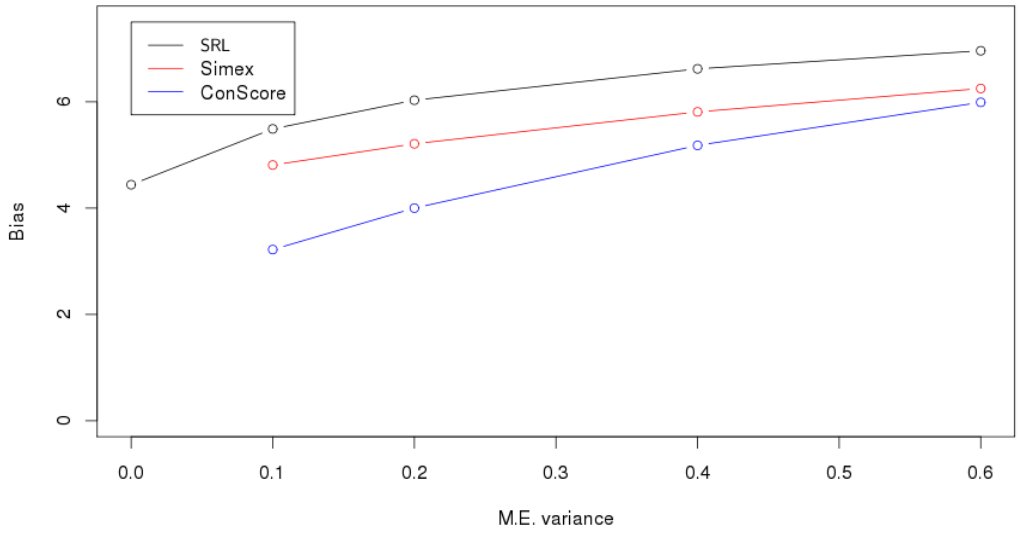

Figure 39: Bias of the square root lasso estimates and its corrections.

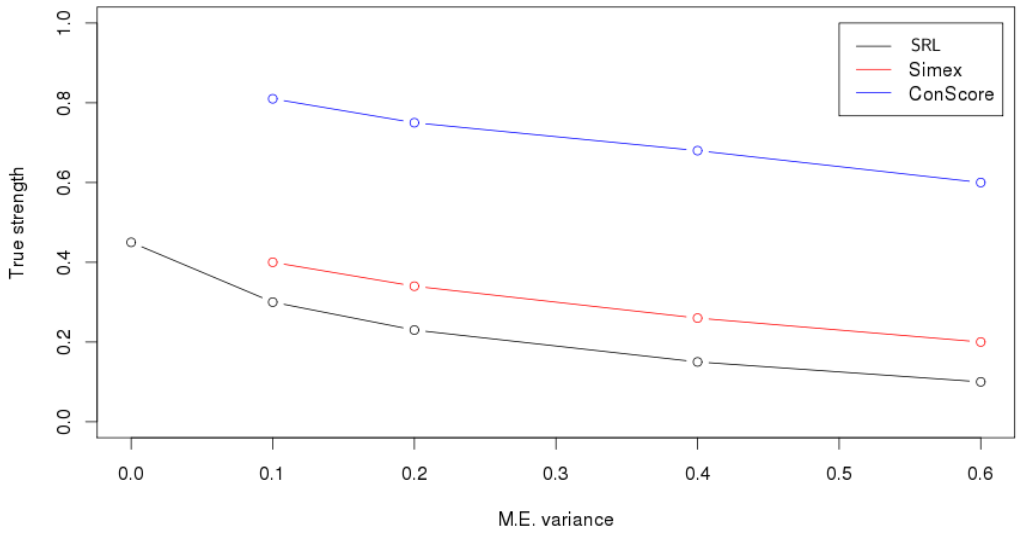

Figure 40: Strength at true locations for the square root lasso estimates and its corrections.

The plots of bias/true proportion confirm that the sparse conditional score performs much better than the naive and SIMEX correction for the SRL. Nev-

ertheless, at the sparsity level chosen for these Monte Carlo simulations, the SRL estimates compare poorly with the other methods.

The stand out methods from the PED correction appear to be SIMEX and the PED conditional score (using all the covariates as opposed to the Sparse CS, where the reduced problem was used). The SIMEX again performs best in terms of the true positives and the strength at the true locations. On the other hand though, the bias and the number of false positives are particularly bad. The conditional score methods give better performance in these areas but do not give the same level of correction at the true locations.

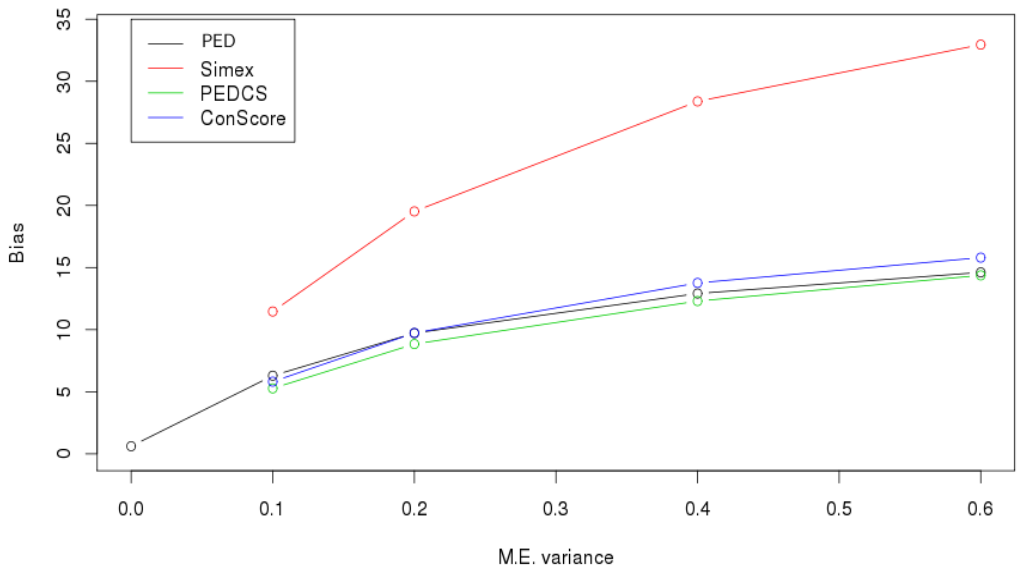

Figure 41: Bias of the PED estimates and its corrections.

It is worth noting that while the reduced, post-sparse version of the conditional score (Sparse CS in tables, using the covariate selection from the naive estimate, see section 4.4) does not perform best in any category, it probably gives the best compromise overall, falling as it does between SIMEX and PED

CS in its performance. It also has the advantage of being much less computationally expensive.

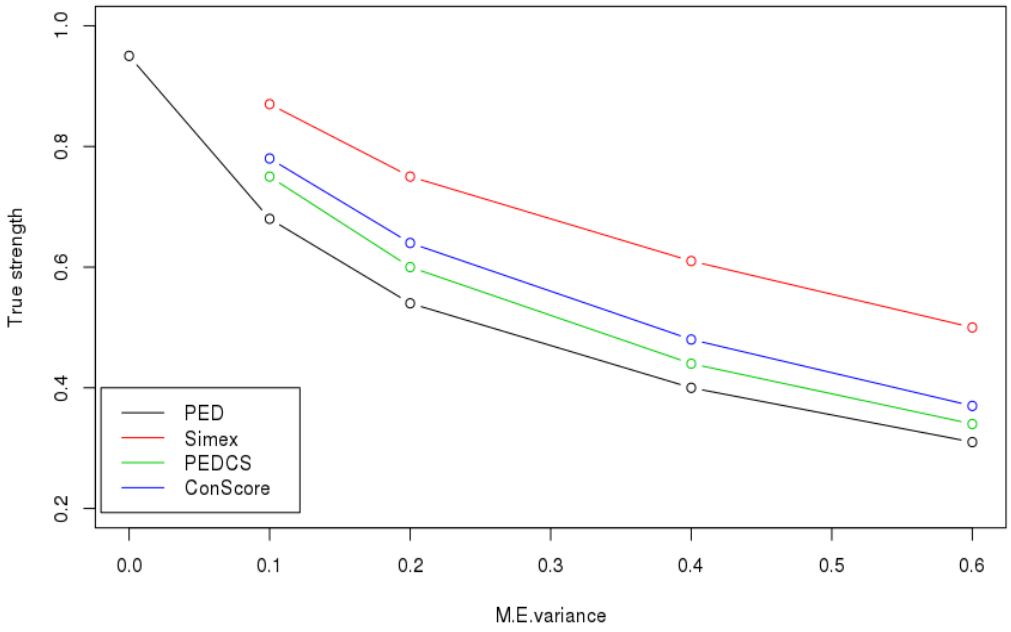

Figure 42: Strength at true locations for the PED estimates and its corrections.

On the note of computational expense, it is important that we consider the practicality of the correction methods alongside their performance. The SIMEX estimates consistently perform well in terms of the correction of the true covariates, yet it is also the most computationally expensive. For each time point it requires us to calculate at least 500 naive estimates depending on the number of iterations and values of the scaling factor ζ that we use. Then we have to compute a quadratic model for each of the covariates to give the corrected estimate. Likewise, the optimisation required in the conditional score methods mean that for larger dimensional problems, it will become more difficult to find an optimal solution.

The calculation of the naive estimates over the 250 Monte Carlo simulations took around one and a half days in *R*. It is easy to see that the SIMEX

estimates will take many times that period of calculation. The calculation time can be reduced by using the same penalty level from the naive estimates across the SIMEX stage, however even then the calculation of the SIMEX estimates in these simulations for all the sparse methods at each of four levels of measurement-error took a little over two weeks in *R*.

σ_u^2	Method	Bias	True +	False +	prop true
0	PED	0.61 (0.63)	4.18 (0.75)	4.22 (7.51)	0.95 (0.08)
0.1	Naive	6.29 (2.49)	3.87 (0.84)	22.36 (6.80)	0.68 (0.08)
	SIMEX	11.46 (4.10)	**4.20 (0.79)**	92.51 (16.03)	**0.87 (0.13)**
	PEDCS	**5.27 (2.71)**	3.69 (0.90)	**7.34 (6.07)**	0.75 (0.12)
	SparseCS	5.81 (2.50)	3.87 (0.84)	22.36 (6.80)	0.78 (0.09)
0.2	Naive	9.72 (3.70)	3.81 (0.84)	29.10 (5.73)	0.54 (0.08)
	SIMEX	19.52 (6.96)	**4.25 (0.80)**	142.48 (22.76)	**0.75 (0.13)**
	PEDCS	**8.84 (3.11)**	3.51 (0.95)	**13.06 (5.37)**	0.60 (0.12)
	SparseCS	9.73 (3.80)	3.81 (0.84)	29.10 (5.73)	0.64 (0.09)
0.4	Naive	12.91 (4.75)	3.73 (0.87)	35.84 (5.16)	0.40 (0.07)
	SIMEX	28.38 (10.27)	**4.35 (0.70)**	200.07 (32.11)	**0.61 (0.13)**
	PEDCS	**12.30 (4.04)**	3.37 (0.92)	**19.16 (5.05)**	0.44 (0.10)
	SparseCS	13.76 (4.97)	3.73 (0.87)	35.84 (5.16)	0.48 (0.09)
0.6	Naive	14.61 (5.30)	3.55 (0.84)	40.34 (5.38)	0.31 (0.06)
	SIMEX	32.95 (12.04)	**4.35 (0.74)**	231.94 (37.62)	**0.50 (0.12)**
	PEDCS	**14.37 (4.27)**	3.17 (0.97)	**22.88 (4.19)**	0.34 (0.08)
	SparseCS	15.80 (5.48)	3.55 (0.84)	40.34 (5.38)	0.37 (0.08)

Table 18: Results summary for PED on 250 Monte Carlo simulations.

Therefore, for higher dimensional problems, it may be preferable to use a reduced version of the problem by using the selection from the naive estimate for reasons of computational expediency.

7.2 MULTIPLICATIVE

For multiplicative measurement-error the conditional score method is no longer appropriate and therefore our comparisons are restricted to SIMEX and the corrected elastic net. Furthermore, the corrected elastic net is obviously only suitable for the lasso, elastic net and ridge regression. Tables 19 and 20 compare the naive, SIMEX and corrected estimates for the lasso and elastic net respectively. For the square root lasso and PED, refer to Tables 10 and 11 in chapter 5 for the naive and SIMEX estimates respectively.

σ_u^2	Method	Bias	True +	False +	Prop-tr
0	Lasso	0.78 (0.27)	4.83 (0.43)	25.52 (13.86)	0.95 (0.03)
0.01	Naive	1.33 (0.53)	**4.76 (0.51)**	25.59 (15.37)	0.91 (0.03)
	SIMEX	1.35 (0.72)	4.66 (0.57)	4.92 (6.32)	0.89 (0.04)
	CorLass	**0.49 (0.23)**	4.67 (0.58)	**1.85 (2.38)**	**0.99 (0.02)**
0.04	Naive	2.32 (1.22)	**4.57 (0.66)**	24.56 (16.23)	0.83 (0.05)
	SIMEX	2.93 (2.06)	4.51 (0.68)	9.74 (9.04)	0.80 (0.06)
	CorLass	**1.46 (1.11)**	4.44 (0.71)	**4.14 (4.04)**	**0.95 (0.06)**
0.16	Naive	**4.47 (2.13)**	**4.20 (0.82)**	23.87 (17.63)	0.65 (0.07)
	SIMEX	5.98 (3.67)	4.18 (0.81)	19.60 (10.68)	0.69 (0.10)
	CorLass	5.409 (3.36)	4.12 (0.84)	**14.34 (8.76)**	**0.95 (0.24)**
0.25	Naive	**5.41 (2.68)**	**4.12 (0.82)**	24.30 (19.69)	0.57 (0.08)
	SIMEX	7.10 (5.32)	4.03 (0.84)	21.20 (12.64)	0.70 (0.12)
	CorLass	6.81 (3.27)	4.01 (0.85)	**17.80 (8.64)**	**0.85 (0.21)**

Table 19: Results summary for lasso with multiplicative error.

σ_u^2	Method	Bias	True +	False +	Prop-tr
0	EN	1.18 (0.34)	4.84 (0.43)	37.99 (13.53)	0.93 (0.03)
0.01	Naive	1.96 (0.69)	**4.75 (0.51)**	38.34 (14.73)	0.88 (0.03)
	SIMEX	1.74 (0.77)	4.68 (0.56)	7.07 (6.40)	0.87 (0.04)
	CorEN	**0.57 (0.29)**	4.64 (0.60)	**2.40 (3.92)**	**0.98 (0.02)**
0.04	Naive	3.33 (1.57)	**4.58 (0.64)**	39.15 (18.60)	0.80 (0.05)
	SIMEX	3.89 (2.33)	4.51 (0.68)	15.89 (11.22)	0.78 (0.06)
	CorEN	**2.25 (1.19)**	4.50 (0.68)	**9.49 (4.28)**	**0.94 (0.05)**
0.16	Naive	**5.72 (2.62)**	**4.26 (0.80)**	35.77 (19.79)	0.61 (0.07)
	SIMEX	7.57 (4.532)	4.17 (0.82)	27.92 (12.41)	0.66 (0.10)
	CorEN	6.33 (3.17)	4.124 (0.83)	**19.63 (8.54)**	**0.91 (0.20)**
0.25	Naive	**6.52 (3.10)**	**4.13 (0.85)**	33.50 (21.55)	0.52 (0.08)
	SIMEX	8.53 (5.50)	4.04 (0.85)	28.87 (13.53)	0.64 (0.13)
	CorEN	8.565 (4.39)	4.04 (0.85)	**25.02 (9.40)**	**0.86 (0.25)**

Table 20: Results summary for elastic net with multiplicative error.

From Tables 19-20 the corrected elastic net/lasso consistently performs the best in the Monte Carlo simulations. The number of true positives is fairly consistent for the naive, SIMEX and corrected estimates with the naive estimate being marginally higher on average. Otherwise, the corrected estimate has a lower average number of false positives and an increase in the proportion of true activity detected, albeit along with greater variance as discussed in section 6.5.2.

The multiplicative measurement-error seems to have had a lesser effect compared to the additive data. It seems that the measurement-error level used is relatively smaller for the multiplicative case. Due to the nature of the log-normal distribution we get more extreme values for the multiplicative measurement-error, but a greater concentration of the error values are around smaller values.

In Fig. 43 we have plotted some random normal predictors, X, along with examples of additive ($W_1 = X + U_1$) and multiplicative ($W_2 = X * U_2$) error. In these 50 points, the variances of W_1 and W_2 are comparable, however the sum of the absolute difference from X is much larger for the additive case (roughly 27 compared with 12). As we can see from Fig. 43 the multiplicative error contaminated green points are generally much closer to the original, black points (with some notable extreme exceptions). In fact for 35 out of the 50 points, W_2 is closer to X than W_1.

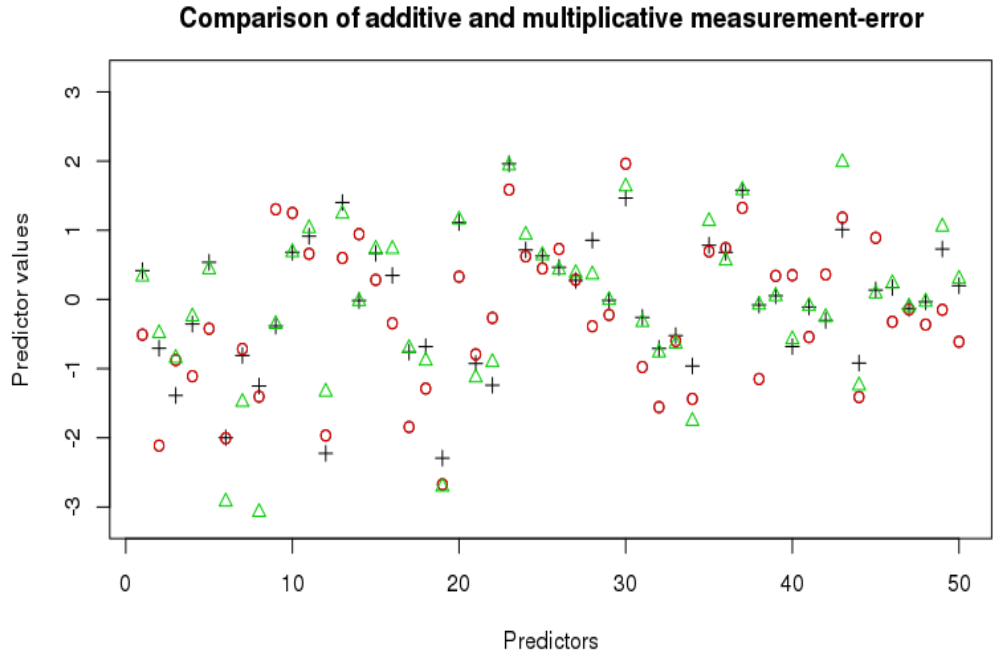

Figure 43: Random $X \sim \mathrm{N}(0, 1)$ (black), additive measurement-error $W_1 = X + U_1$ where $U_1 \sim \mathrm{N}(0, 0.5)$ (red), multiplicative measurement-error $W_2 = X * U_2$ where $U_2 \sim \mathrm{lnN}(0, 0.2)$ (green).

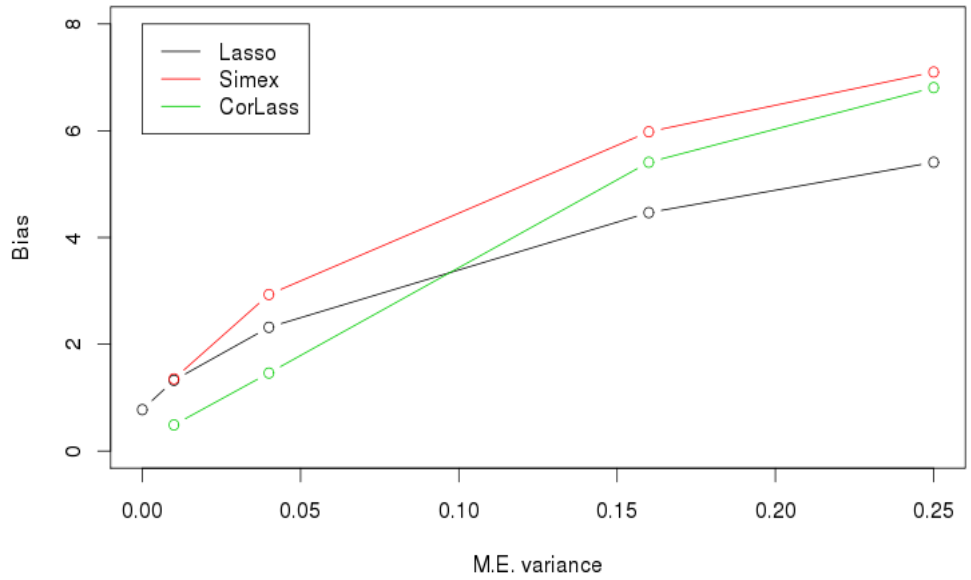

Figure 44: Bias of the lasso estimates and its corrections for multiplicative error.

Fig. 44 shows that the bias follows a similar pattern to the additive measurement-error case, with the corrected lasso displaying the smallest bias for the lower measurement-error levels. At the true locations, the corrected lasso is consistently stronger than the naive lasso and SIMEX estimates. This is particularly evident in Fig. 45. The naive estimate is fairly good at the true source locations for low measurement-error and consequently the SIMEX estimate is only an improvement when we increase the error. However even then the corrected lasso performs much better, although there is the aforementioned issues with the increasing variance of the estimates.

As with the additive case, the elastic net estimate is very similar to the lasso and the grouping property does not seem to improve the performance for uncorrelated covariates.

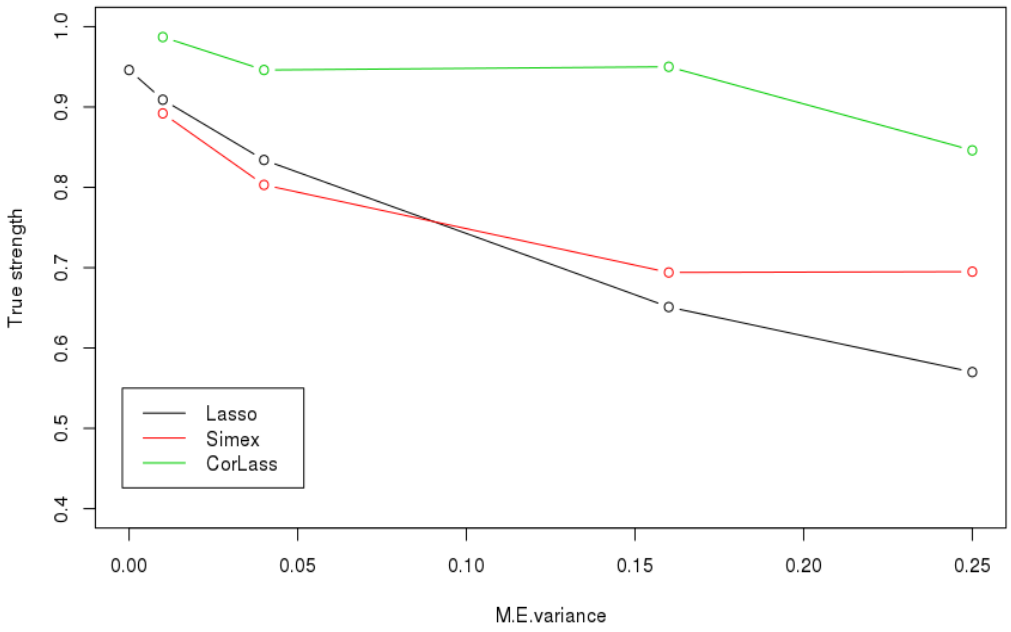

Figure 45: Strength at true locations for the lasso estimates and its corrections for multiplicative error.

7.3 ADDITIVE MEG SIMULATIONS

Thus far the measurement-error correction methods have shown good performance in Monte Carlo studies for both additive and multiplicative error. In each of these cases however, the simulations have dealt with independent, non-correlated parameters. Therefore, the highly structured nature of the leadfield matrix and the naturally correlated covariates in MEG data will present a much more challenging environment for the measurement-error methods. Furthermore, the aforementioned properties will be reflected in the type of measurement-error we would expect in MEG experiments. It is likely that assuming constant error variance across all locations/orientations represented in the leadfield matrix is far too simplistic. Instead we will have to deal

with different error variances for each location as well as potentially correlated measurement-error for nearby locations.

Before moving to some real data, the measurement-error methods are applied to some single slice MEG simulations. Starting with additive error, a single source of strength 5 is simulated for 100 time points. Noise is added to the data from a normal distribution with mean 0 and variance 3 (the variance level was chosen to represent data where the signal-to-noise ratio had been increased through trial averaging). Treating the calculated leadfield matrix as the true forward model, error is added from a multivariate normal distribution with mean vector $\mathbf{0}_p$. In order for the variance to differ according to the location, the error covariance matrix used is a diagonal matrix where each element of the lead diagonal is the standard deviation of the relative column of the leadfield matrix. For the minimum norm and sparse regression methods, an estimate is obtained under no measurement (using the original leadfield) along with the naive estimate (using the contaminated leadfield) and the corresponding measurement-error corrected estimate for the conditional score, SIMEX (with both a linear and quadratic extrapolation function) and corrected elastic net (including the special cases of the minimum norm and lasso). Under the single slice simulation we have $n = 270$ observations and $p = 556$ covariates. The plots that follow give the mean estimates for the measurement-error free case, the naive estimate and the estimate after employing the measurement-error methods.

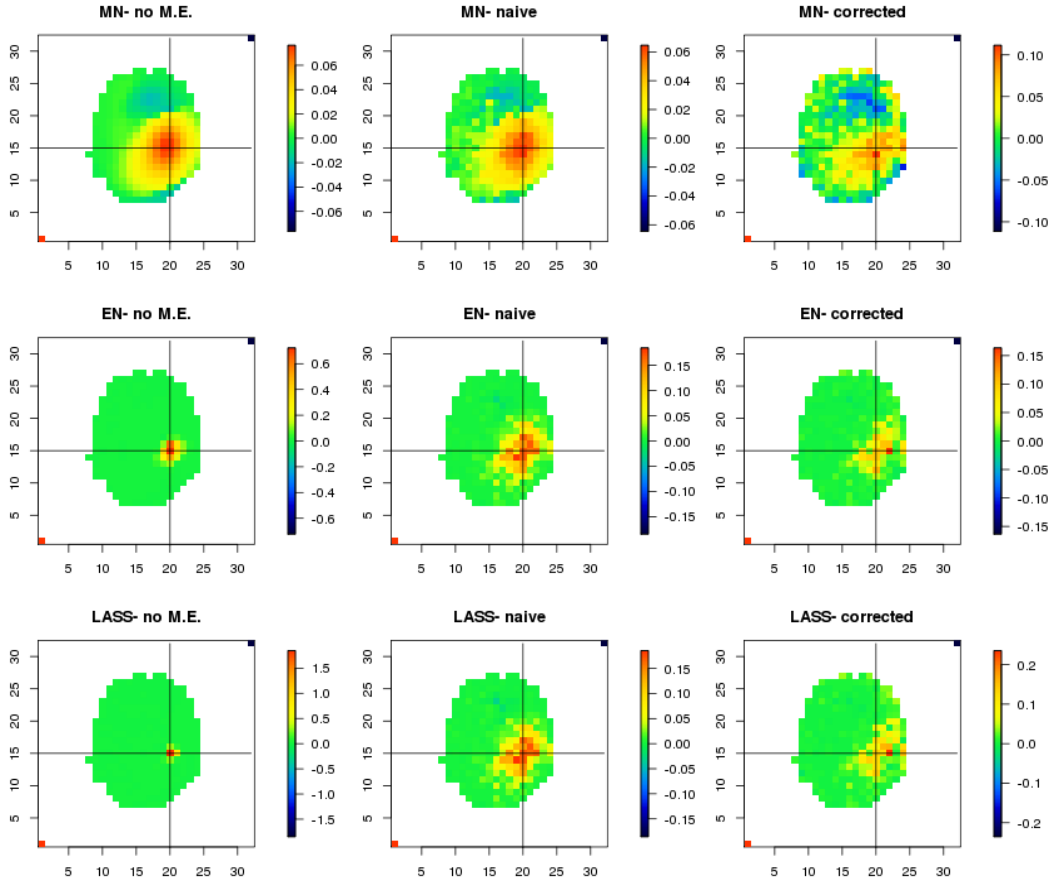

Figure 46: Estimates under no measurement-error, naive estimates and corrected estimates as detailed in chapter 6 (l-r) for minimum norm, lasso and elastic net (top to bottom). Additive measurement-error.

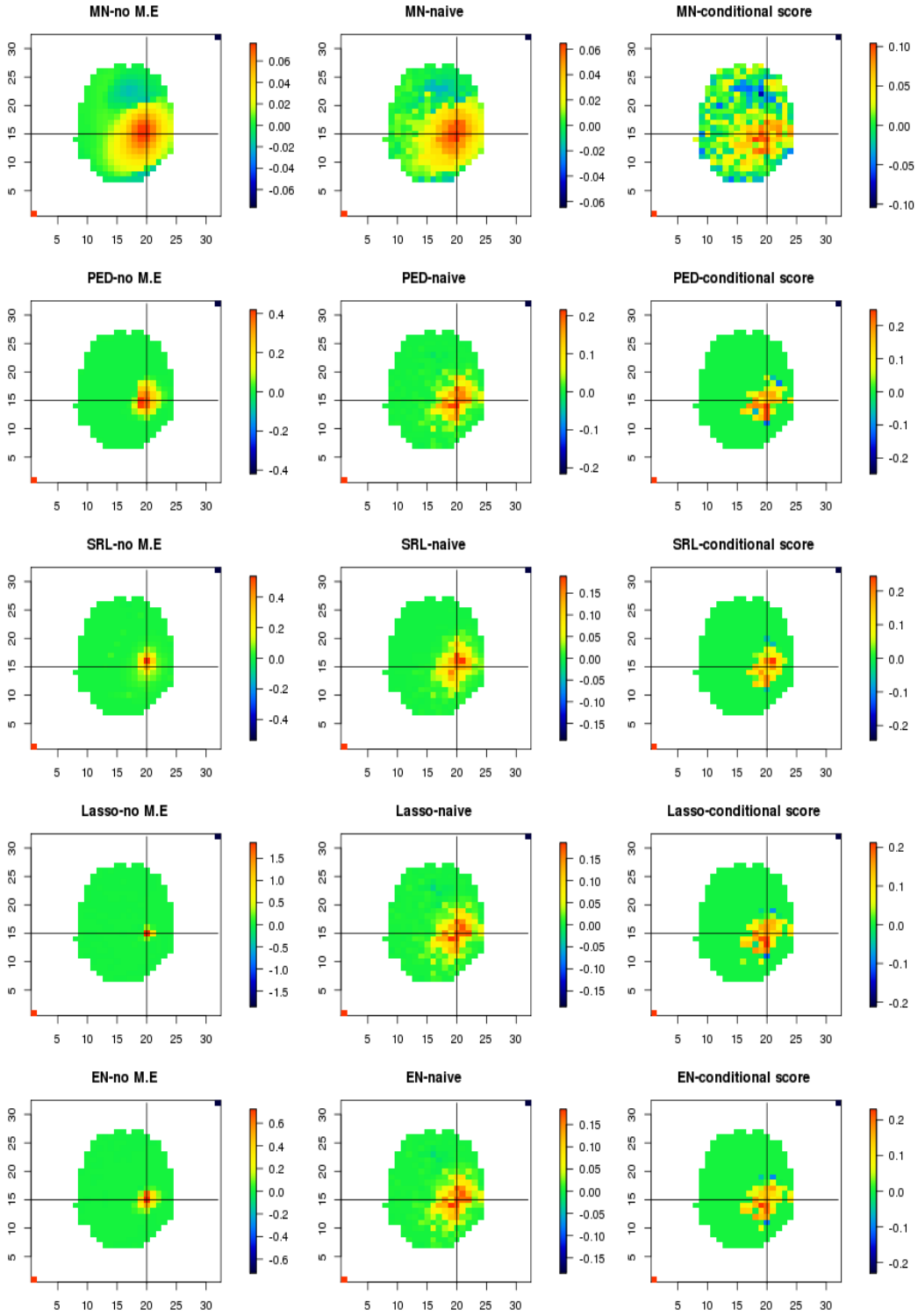

Figure 47: Estimates under no measurement-error, naive estimates and conditional score estimates (l-r). Additive measurement-error.

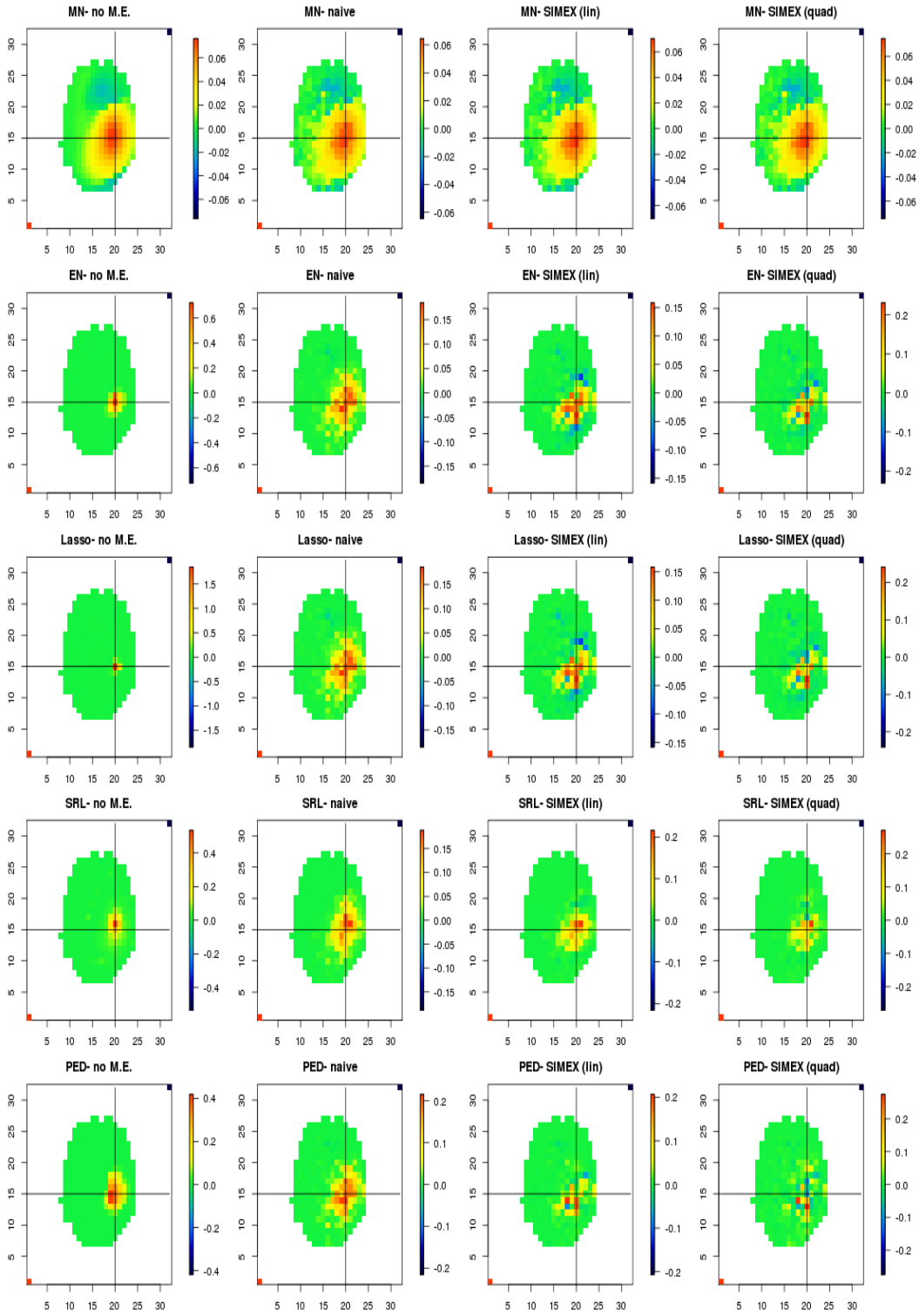

Figure 48: Estimates under no measurement-error, naive estimates and SIMEX estimates using a linear and quadratic extrapolation function (l-r). Additive measurement-error.

The corrected estimates in Fig. 46 show that the minimum norm remains relatively unchanged under the additive error. The corrected minimum norm method is much noisier than either the estimate with the true leadfield or the naive estimate, although interestingly the strongest pixel is approximately in the correct location. The main difference between the naive and corrected estimates is the placement of a stronger peak one pixel away from the true source location in the corrected case. The elastic net and lasso methods display much more deterioration in their naive counterparts as the placement of activity is much wider. In these cases, the corrected method seems to place a single stronger source among the weaker activity rather than the group of darker pixels seen in the naive estimate. Unfortunately, the placement of this source is more peripheral than the naive estimate.

The conditional score method seems to give stronger sources than the naive estimate as displayed in Fig. 47. For the minimum norm case the conditional score estimate is noisy, but the peak is almost twice the strength of the naive equivalent and is placed fairly sensibly. For the sparse methods, where only the naive selected locations are used, the conditional score rescales and re-concentrates the estimates a little. This second point is demonstrated by the assignment of a small number of strong pixels in the conditional score estimate rather than the larger clump of strong sources in the naive estimates.

Finally, the SIMEX estimates from Fig. 48 seem to show the biggest variation in performance. The minimum norm SIMEX method gives a slightly rescaled estimate for both linear and quadratic extrapolations, but is essentially unchanged otherwise. Given the gentility of the change between the original and naive estimates, this is not necessarily a bad thing. The SIMEX estimates for the sparse methods are more varied. For the lasso, elastic net, and to a lesser

extent PED, SIMEX removes a number of the weaker yellow pixels but also sees a number of negative sources. The linear extrapolation for the lasso and elastic net in particular places strong negative sources. In these cases, one of the components of the pixels source has evidently been over-corrected resulting in a sign change. As a result, the estimates seem much noisier than those in Fig. 47. The quadratic extrapolation seems more appropriate in these cases and we end up with much sharper peaks than in the conditional score estimates, but the placement of the main source peak is a little lower than it should be. The square root lasso gives the best results from the SIMEX estimates. A number of the false positives are reduced, the strength of the peak is increased and its size is more localised.

7.4 MULTIPLICATIVE MEG SIMULATIONS

Moving on to multiplicative measurement-error, the true signal remains unchanged from the additive case and our new model error is simulated from a log-normal distribution. Once again, we want the measurement-error variance to vary according to the location of the corresponding parameter in the head, therefore each row, $\mathbf{U}_i$, $i = 1, 2, ..., n$, of the measurement-error matrix, $\mathbf{U}$, is simulated from $\ln\mathrm{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_u)$, where $\boldsymbol{\Sigma}_u$ is the $p \times p$ diagonal matrix with the elements of the lead diagonal represented by $\boldsymbol{\Sigma}_{u,(i,i)} = \log(\mathrm{Var}(\mathbf{L}_{,i}))/4$. i.e. each diagonal element is proportional to the natural log of the variance of the corresponding column in the leadfield matrix. The multiplicative forms of SIMEX and the corrected elastic net are performed on the data, whereas the conditional score is no longer appropriate. Again we compare the measurement-error method with the naive and uncontaminated estimates.

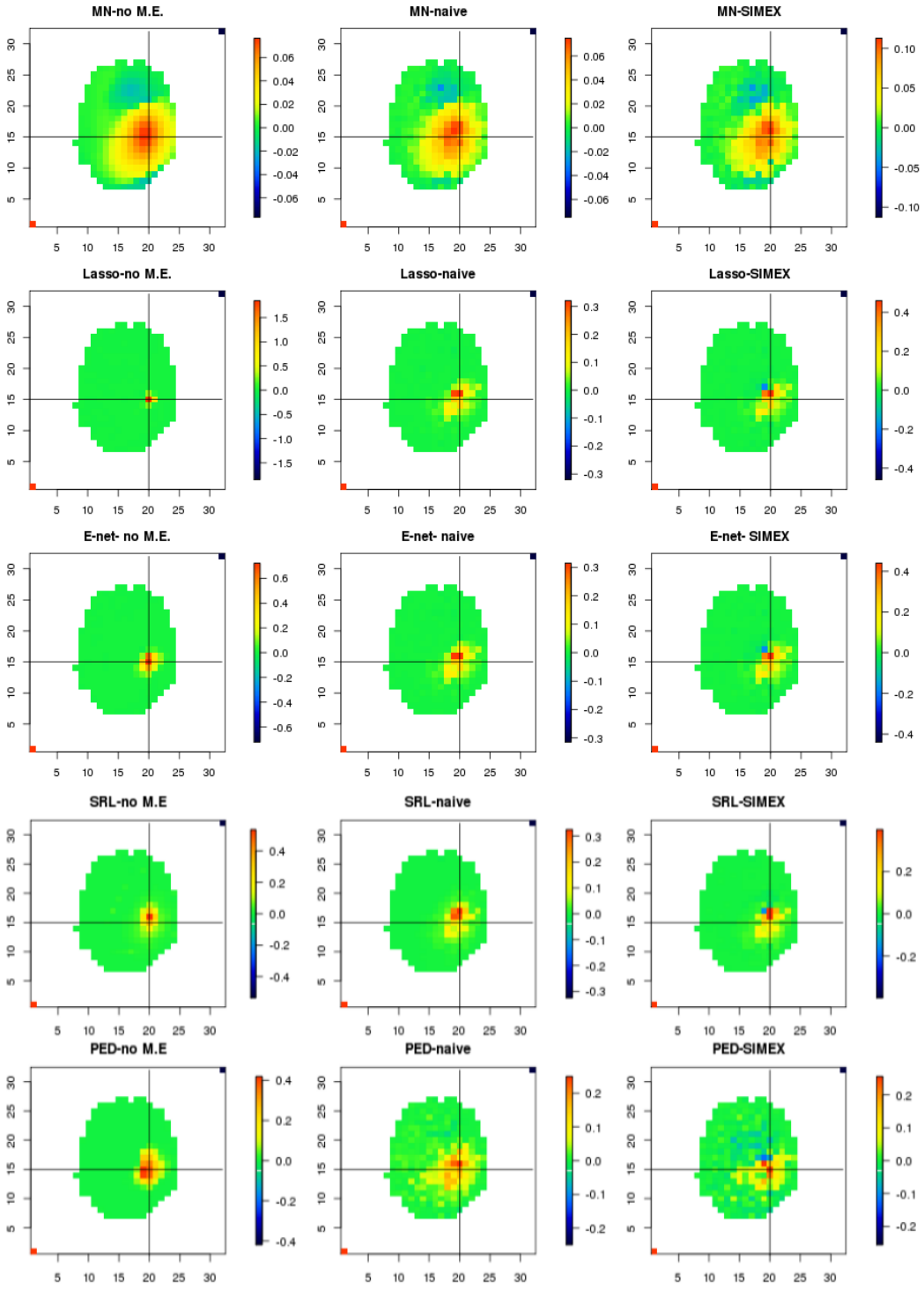

Figure 49: Estimates under no measurement-error, naive estimates and multiplicative SIMEX estimates using linear extrapolation (l-r).

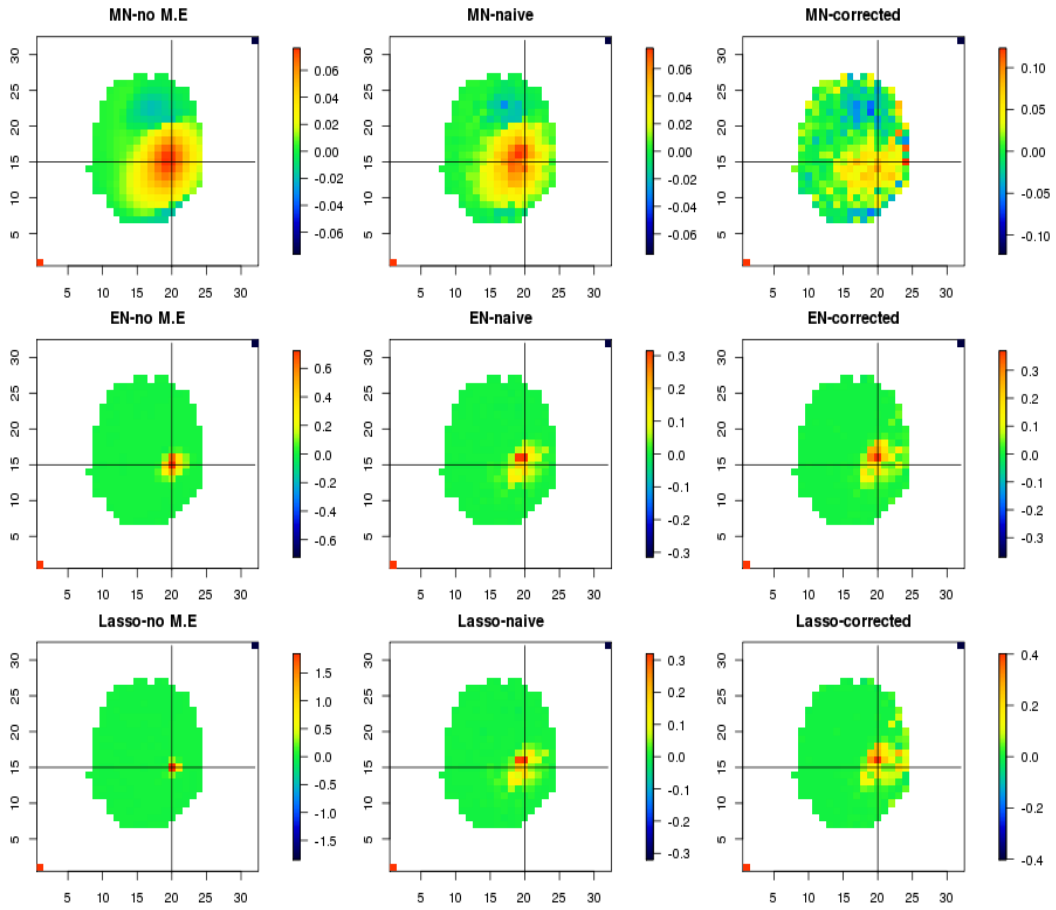

Figure 50: Estimates under no measurement-error, naive estimates and multiplicative corrected elastic net estimates (l-r).

From Fig. 49 the SIMEX estimate for the minimum norm has undergone a rescaling, but otherwise shows little change from the naive. Again though, as with the additive error, the naive estimate is so close to the standard minimum norm without measurement-error that it is questionable whether a correction is needed. The sparse regression methods display a greater need for a correction as the placement of activity disperses under their naive implementations. PED shows the most dramatic change as the number of obvious false positives increases drastically. Some of these points are removed after SIMEX has been performed, however the most significant change that SIMEX achieved is the placement of a source at the true location. The naive estimates for the lasso,

elastic net and square root lasso are very similar and this resemblance is retained with the SIMEX correction. In each of these cases the SIMEX estimate has a stronger peak and has condensed the main peak of activity to one (or two in the case of SRL) pixel(s) adjacent to the true source.

Of course, when we move into real data the ground truth is unknown, therefore the 'true location' is no longer useful as a measure of the performance of the sparse methods. We can use a numerical metric like the mean square error to measure how close the fitted values are to the recorded data, however this will not give us information about the spatial layout of the estimated sources. Therefore the location of the distributed sources can only be assessed in comparison to other methods and our knowledge about the functionality of brain regions.

The corrected elastic net shows similar properties to SIMEX under multiplicative measurement-error as it places a single stronger peak one pixel above the true location in both the elastic net and lasso. However there is also an introduction of a number of false positives on the right hand edge of the slice. This is particularly evident in the minimum norm correction where the strongest sources have been dragged to the edge. The columns of the leadfield associated with these locations have some of the highest variances due to their proximity to the sensors. Furthermore, given the chosen distribution of the measurement-error, the correction at these points will be larger. Under multiplicative measurement-error, the minimum norm (ridge) case of the corrected elastic net does not seem appropriate, especially given that the introduction of measurement-error seems to have very little impact on the performance.

Again the effect of the multiplicative error appears to much less dramatic than the additive noise. It is possible that the level of multiplicative error used

works out to be lower than that of the additive measurement-error and from Fig. 51 the smaller values in the leadfield show less divergence in the multiplicative error. However for larger values in the leadfield, the multiplicative error works out be larger. Due to the nature of multiplicative error, further increases of the error variance would result in some extreme values of error for the larger elements in the leadfield.

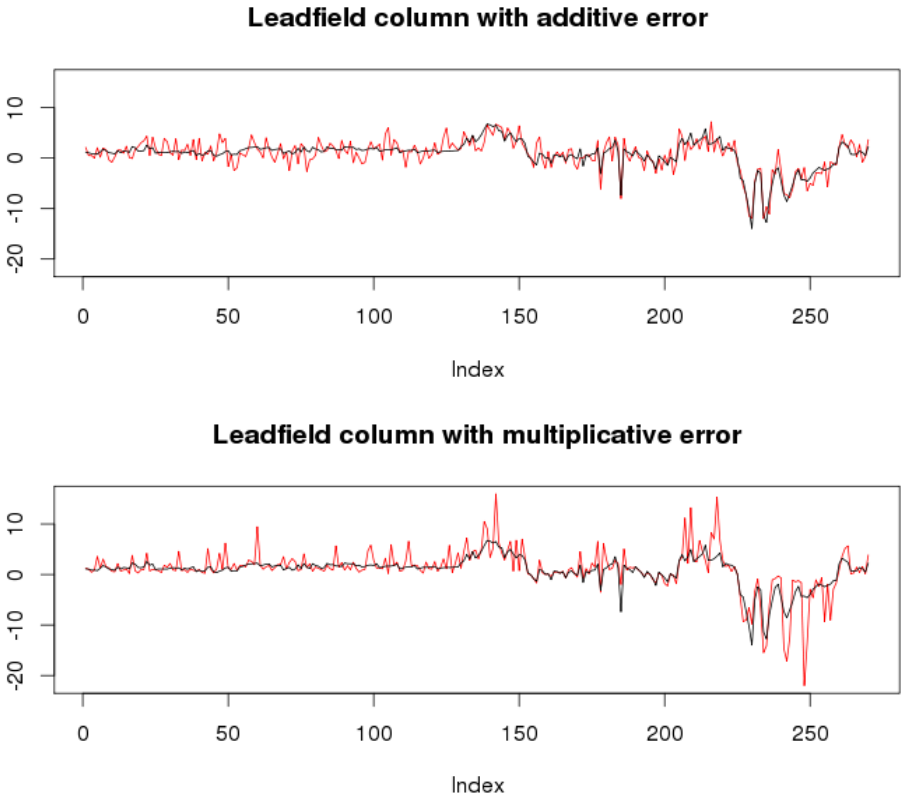

Figure 51: Measurement-error contaminated leadfield for additive (top) and multiplicative (bottom) error. Distributions of the error described in 7.3 and 7.4 respectively.

7.5 CONCLUSIONS

In the presence of measurement-error, there is an increase in the number and/or magnitude of false positives in the estimates chosen by the sparse regression methods. In single source MEG simulations the estimates were shown to lose some spatial resolution and wider areas of activity were chosen over the more precise estimates when there is no measurement-error present. This was particularly evident in the PED solutions, which became increasingly noisy with the introduction of error into the model. The lack of precision in the minimum norm means that the naive estimate shows very little deviation from the error free estimate, even under substantial measurement-error. Because of this, the application of measurement-error correction methods to the minimum norm in MEG appears to achieve very little other than rescaling. Therefore, since the method appears to be very robust to measurement-error, it seems fair to conclude that measurement-error correction for the minimum norm is largely unnecessary.

On the other hand, correction methods seem much more appropriate for the sparse regression methods since they are more affected by measurement-error. In simulations involving uncorrelated variables the conditional score (for additive error), SIMEX and corrected elastic net all show an improvement over the naive estimate. The latter is particularly effective for multiplicative error but is obviously restricted to use with the lasso, elastic net and ridge (at a stretch). The performance of the measurement-error corrections in the highly correlated, more structured MEG data is much more modest. In these circumstances the measurement-error methods give a slightly more refined peak with a stronger source, we also see some adjustment to some of the false positives.

Therefore it seems like SIMEX, despite the computational issues, is the most appropriate method as it performs well in the simulations and can be used for all the sparse methods. The main issue as we move into real data is that the dimensions dramatically increase. However, using the selection from the naive estimate and only performing SIMEX on the selected parameters can greatly alleviate some of the computational burden. The corrected elastic net often displayed better performance than SIMEX in the simulations, but is restricted in only being valid for the lasso and elastic net.

If we now consider the use of measurement-error methods for real MEG data, from the MEG simulations in this chapter it seems reasonable to assume that multiplicative error is more appropriate than additive error. Due to the structured nature of the leadfield matrix, an incorrect scaling in the modelling stage of the forward calculations seems a likely source of error. Furthermore, it is a reasonable assumption that the error variance will depend on the location of the parameter in the head. Therefore, for the real data we will continue to use the variance of the leadfield columns to help infer the measurement-error.

8

MEASUREMENT-ERROR IN REAL MEG DATA

Real data brings extra complexity to the measurement-error methods. Along with the increasing dimensions of the data in MEG we also have to determine the distribution that the measurement-error follows. Indeed, considering the form that the measurement-error takes is important to the performance of the corrections. The methods thus far have assumed that the measurement-error variance is known, or at least able to be estimated. However, in some cases knowledge of the data will have to be used to inform the choice of the estimated measurement-error level.

In the applications to the real MEG data, a multiplicative model is a sensible choice for the measurement-error, i.e. $\ln\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$. The multiplicative model allows us to introduce error into the leadfield that is proportional to the size of the original elements of the matrix. Since the size of the elements of the leadfield matrix vary massively according to their location, the addition of uniformly variable errors makes little sense. The log-covariance matrix $\boldsymbol{\Sigma}_u$ was chosen to be a diagonal matrix with lead diagonal elements, $\Sigma_{u,(i,i)} = \log(\mathrm{Var}(\mathbf{L}_{,i}) + 1)/5$ in a similar way to the previous MEG simulations (the addition of the $+1$ was added to ensure the column variances of $\mathbf{L}$ were > 1). From the analysis in section 2.4.2 we can see that the estimates from the sparse methods are

fairly smooth. This suggests that any error in the leadfield is fairly moderate. Increasing the values in $\mathbf{\Sigma}_u$ significantly leads to some rather extreme values when sampling from the log-normal distribution. Therefore, the matrix $\mathbf{\Sigma}_u$ was chosen to be able to account for fairly modest measurement-error in our real data.

The main computational challenge for the measurement-error methods when moving to real data is the increase in dimensions involved in the data. For the five slices chosen for source localisation in our data there are 2734 potential locations/orientations of interest. Under ideal circumstances with no measurement-error, dramatic increases in the number of covariates will result in slower estimation. However, for the measurement-error methods the issues of estimation time will be more acute. For example in our SIMEX estimation we use 50 repetitions at each of 4 levels of increasing measurement-error, meaning that we are required to make 200 estimates in addition to the original naive estimate. For that reason, the corrections for the sparse methods will be performed using the chosen locations from the original estimates. The multiplicative forms of the corrected elastic net and SIMEX methods are applied to the data previously used in chapter 2. Our analysis is restricted to the primary period of activity in the data between time points 76 and 200.

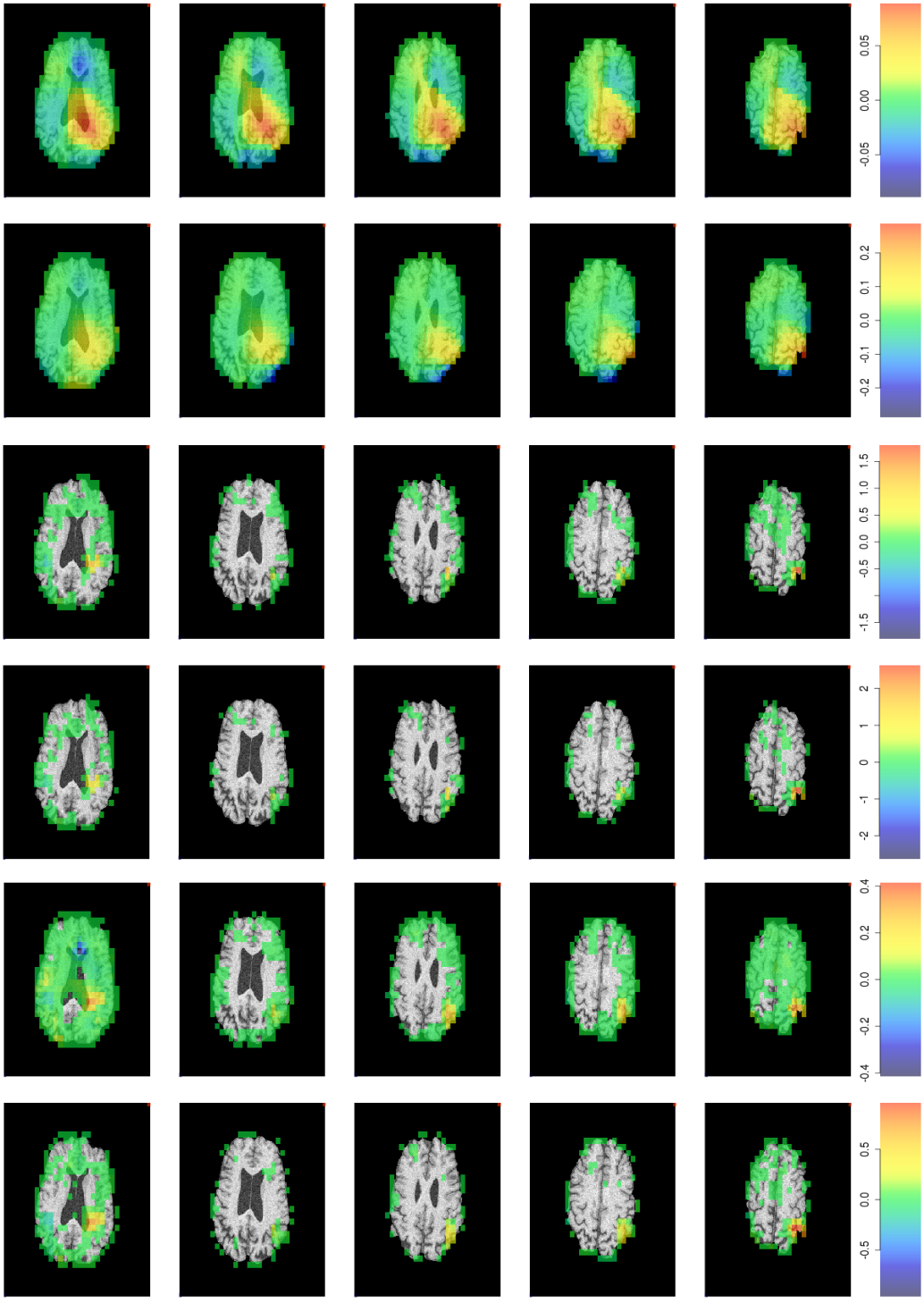

Figure 52: Top to bottom; naive MN, corrected MN, naive lasso, corrected lasso, naive EN, corrected EN. Columns from L-R lowest slice to highest slice in head.

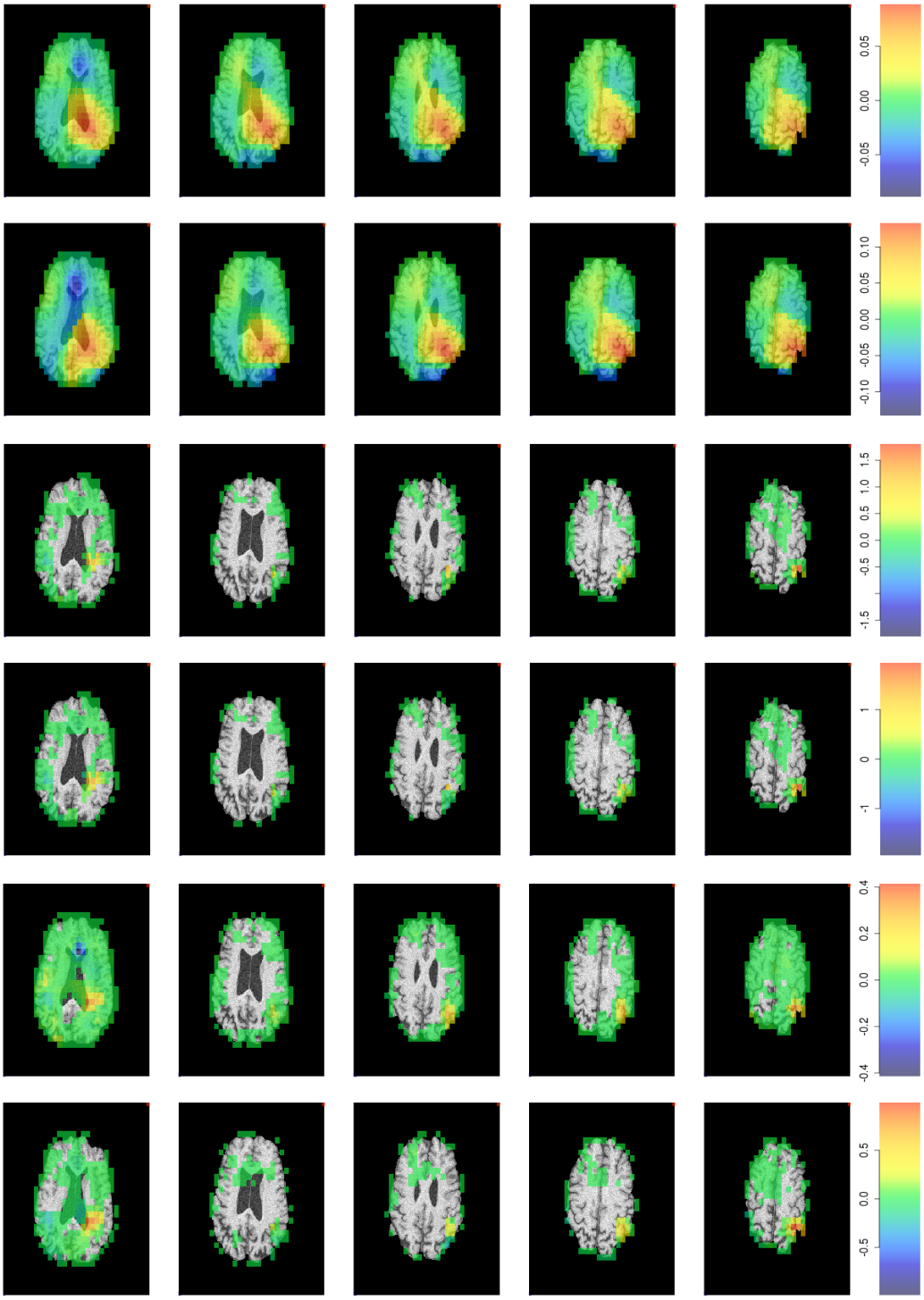

Figure 53: Top to bottom; naive MN, SIMEX MN, naive lasso, SIMEX lasso, naive EN, SIMEX EN. Columns from L-R lowest slice to highest slice in head.

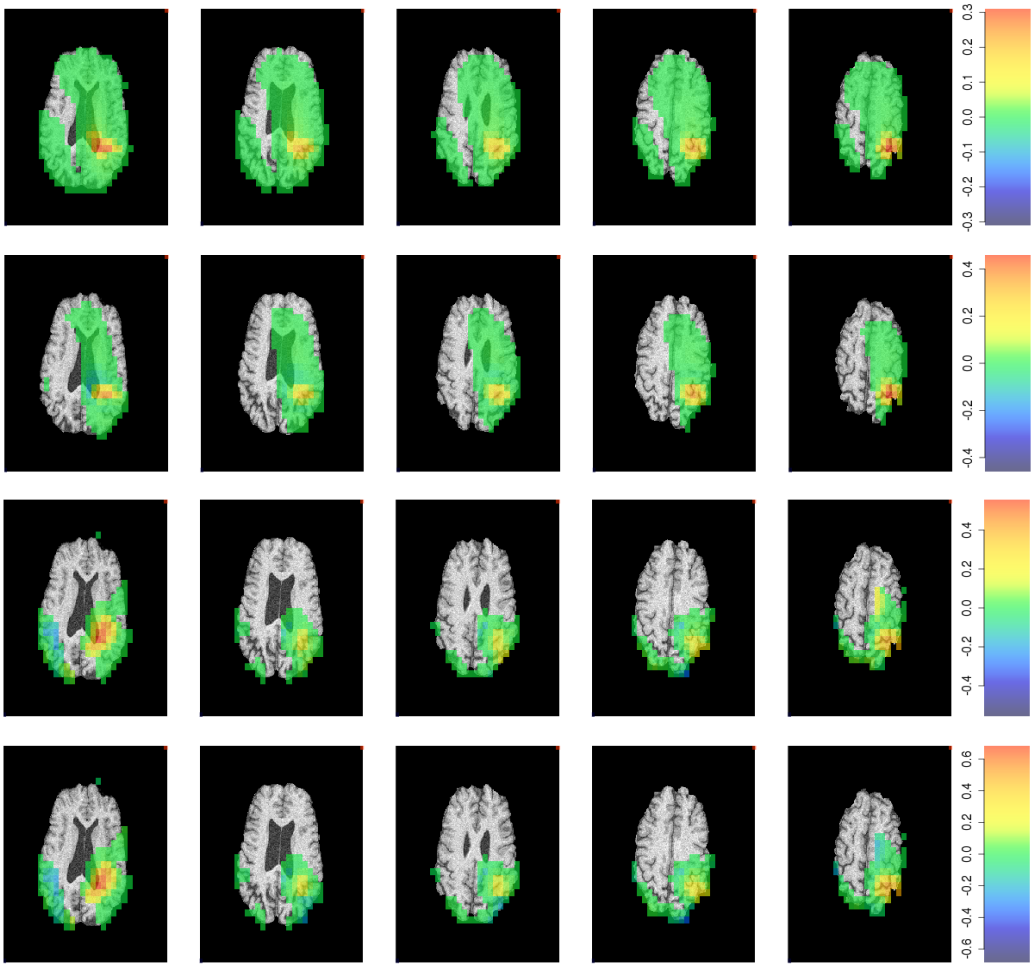

Figure 54: Top to bottom; naive SRL, SIMEX SRL, naive PED, SIMEX PED. Columns from L-R lowest slice to highest slice in head.

The SIMEX estimates for the real data are very much a continuation of the results shown in simulations. As we have come to expect there is a rescaling of parameter estimates and whilst that is the only noticeable difference in the minimum norm estimates, there are a few subtle changes in some of the sparse methods. The SIMEX elastic net estimate has removed the dark blue patch of activity in the lowest slice and has a much more defined peak in the top slice. The SRL SIMEX estimate also places a little less activity in the lowest slice. Additionally, the sparse SIMEX estimates (with perhaps the exception

of PED) assign less noticeable activity in the middle slices, demonstrated by fewer yellow pixels.

The corrected elastic net and lasso display similar minor changes to the SIMEX estimates, with the elastic net again removing the darker blue activity placed in the lowest slice and assigning the primary activity in the upper slice. The most evident change is in the corrected minimum norm which brings the distribution of activity much more into line with the sparse methods. However, the placement of activity at the very edge of the top slice does reinforce the suggestion from the MEG simulations previously in this chapter that the corrected minimum norm is somewhat biased towards the edge locations.

Comparing the two measurement-error approaches, the estimates for the SIMEX and corrected forms of the lasso and elastic net are very similar, although the differences from the naive estimate are very limited. The corrected lasso/elastic net estimates are a little sparser than the SIMEX equivalents, although that is probably to be expected from the nature of the SIMEX method. Although the impact of the measurement-error methods was limited in the simulations, there was some re-distribution of the estimated sources either in a more localised peak or the removal of some peripheral locations from the estimate. The lack of significant change between the measurement-error methods' estimates and the naive does highlight the choice of the measurement-error variance. The measurement-error level was chosen to be moderate in comparison to the size of the leadfield elements, but still a significant size for log-normal error. Therefore, either the original estimates were close to optimal even accounting for measurement-error, or the error level chosen was too conservative. The difficulty of choosing the correct level of measurement-error for the data will be discussed in the next section.

8.1 CONCLUSIONS

MEG data presents a very challenging set of circumstances for traditional methods of statistical estimation. The dimensions involved in the data far exceed the number of observation channels that are feasible, and the spatial nature of the covariate information naturally translates into high correlation between locations in the brain. Therefore, it is important that any methods applied to MEG data are able to deal with high dimensions, the potential for correlation and large amounts of noise. The sparse regression methods generally showed good results when applied to some real MEG data and the placement of activity was consistent with the widely used beamformer and minimum norm methods. The signal leakage found in the minimum norm is certainly not as evident, as the estimates of activity from sparse methods are much more localised. The best performing sparse methods seem to be the square root lasso, and those that have a grouping property, namely the elastic net and PED. These methods seem to go some way to addressing the issues that the standard lasso has in MEG problems. As we discussed in section 2.5 we often get spiky, discontinuous time courses from sparse methods due to treating each time point as an independent observation. This is a particular problem in the case of the lasso due to the spatial instability that results from its treatment of correlated variables (Huang et al., 2006; Zou and Hastie, 2005). The inclusion of a grouping property ensures that some consideration is given to the correlation structure of the variables and seems to give smoother results. Interestingly, the square root lasso seems to perform much better than the standard lasso in this regard despite having the same penalty.

Interpreting the results for real data brings other challenges. Along with the dimensions involved in real MEG data, whereby methods such as SIMEX become increasingly expensive forcing us to work with a reduced parameter space, we also have unknowns that naturally accompany non-simulated data. Determining the levels of measurement-error that are appropriate for a given dataset, or indeed whether there is measurement-error present, represents the primary challenge. This does pose interesting questions over the suitability of even multiplicative measurement-error for MEG. Due to the structural complexity of the leadfield matrix it is very possible that the measurement-error associated with it is also structured in some way. Two of the most likely sources of error are leadfield modelling inaccuracies and movement of the head in the scanner, which results in co-registration errors. Indeed, recent work on physical measures to restrict this type of head movement has been undertaken in order to address a source of model error (Meyer et al., 2017). Consideration of head movement for the framework of measurement-error would require some investigation of the effect of rotational movement on the forward model. Therefore normal or log-normal measurement-error would perhaps be too simple for modelling this situation. Pursuing this line of investigation would also bring the challenge of adapting the existing measurement-error framework and methods to cover a more structured error.

One way of assessing the impact of measurement-error on MEG estimation would be to employ sensitivity analysis. Sensitivity analysis allows us to test how robust our model is to uncertainty and identify which inputs in the model are the most sensitive to error (Chatterjee and Hadi, 2009). From sensitivity analysis we should also be able to make some inference about the suitability of different levels of measurement-error to the data. Such approaches have been

used for a variety of data types where measurement-error is common(Levine, 1985; Willemsen et al., 1991; Greenland, 1996; Agogo et al., 2016).

9

CONCLUSIONS

In this thesis we have focused on bringing together the statistical concepts of sparsity and measurement-error in the context of MEG data. These two areas of investigation have particularly interesting applications to MEG due to the size of the datasets involved and the importance of forward modelling accuracy in the estimation process.

Sparse regression methods are widely used in statistics for situations where the dimensions of the data are large or when a number of the covariates in a model should be set to zero. Nevertheless the use of sparse regression in MEG has been limited. The ℓ_1 penalised lasso, which is known as the minimum current estimate in the MEG literature, has been shown to suffer from a number of issues when it comes to MEG data (Uutela et al., 1999). Chief among these is an insensitivity to the source location which results in a lack of smoothness in the lasso estimate over both spatial and temporal planes. Consequently the MEG community have developed some alternative methods for employing the ℓ_1 norm minimisation whilst retaining smooth solutions. These methods frequently make use of the singular value decomposition of the data. For example VESTAL (Huang et al., 2006, 2014) involves a projection of the minimisation onto a subspace defined by the most important singular vectors of the data,

resulting in smoother time courses (we briefly outline this method at the end of section 2.5). Another approach was taken by Ou et al. (2009), who applied a group norm to the problem. The sources were expressed using a linear combination of orthogonal basis functions over which the ℓ_2 norm was applied. The ℓ_1 norm was then used to give spatial sparsity (Hamalainen et al., 2010).

Similarly, Tian et al. (2012) proposed a method that produces a raw estimate of the source time courses and then refines the estimate with a two-way regularisation. This involves iteratively optimising with a penalty on the spatial coefficients (typically an ℓ_1 norm) and a smoothing penalty of the temporal features. Considerations for including temporal information into MEG inverse modelling extend beyond producing spatially sparse but temporally smooth solutions. It is a reasonable assumption that a sources current state will depend to some extent on its previous evolution, therefore a significant area of MEG modelling looks at reflecting this through statistical models. Two examples of recent statistical approaches that include temporal dependent sources are Yao and Eddy (2014) and Solin et al. (2016). The former assumes that the sources (and therefore the data that depends on the sources) follow a first order Markov process to produce a probabilistic time-varying source model. Solving the inverse problem then involves using importance sampling methods to find the posterior source model. Solin et al. (2016) used Gaussian process priors with separable covariance functions for the spatial and temporal components of the model. This results in spatio-temporal regularised solutions. Both these approaches employ Bayesian methods to give a predictive distribution for a source at a particular location and point in time.

The issues with the lasso in MEG lie in its dealing of correlated covariates, however we found that the sparse methods that give some consideration to the

spatial correlation give much smoother solutions. The elastic net, PED and SRL all give some balance between sparsity and smoothness. In this regard, the elastic net is the most flexible as it allows different mixing ratios between the two penalties.

Studies into measurement-error have largely focused on devising methods of correction under the standard $n > p$ regression framework. The performance of the lasso under measurement-error has been investigated by Loh and Wainwright (2012), and Sørensen et al. (2015), with the latter also giving some discussion to a conditional scores lasso, but generally investigation into the effects of measurement-error in high dimensional data has been limited. We found that, whilst the smooth ridge estimate was largely unaffected by measurement-error in MEG problems, the naive forms of the sparse regression methods suffer from selection issues when measurement-error is introduced. This results in attenuation bias and the introduction of an increasing number of false positives. The PED estimate is particularly affected by false positives.

In regards to the measurement-error correction methods, we found that the conditional score, SIMEX and corrected estimates were an improvement over the naive sparse estimates in simulations with uncorrelated covariates. Our extension of the corrected lasso to the elastic net was found to give performance in line with the results of Sørensen et al. (2015), although the corrected lasso remains the better choice for uncorrelated covariates. Despite the general improvement, the correction capabilities of the measurement-error methods are largely restricted to addressing attenuation bias and are limited when it comes to selection adjustment. The main issue when correcting sparse estimates is that the measurement-error correction methods tend to increase the magnitude of the false positives in addition to the true positives. Thresholding can be used

to remove some of these false positives, but there is also the risk of removing important covariates and further work needs to be done to address the issues with selection under measurement-error.

Whilst the measurement-error methods provide good correction in uncorrelated covariates, the improvement for correlated variables is much less evident. The complexity of MEG data provides a major challenge to the measurement-error methods and the evidence of improvement over the naive estimate in MEG simulations is limited. In real data studies the correction methods rescaled the estimates and some of the deeper placed activity was removed, but the distribution of activity was largely unchanged. This may mean that the original estimates were close to the truth, or the correlation structure of the MEG data may be too challenging for the methods to provide meaningful corrections.

Going forward one of the areas for further study would be looking at more structured, complex forms of measurement-error. As we mentioned in section 8.1 it is likely that the key sources of error in the MEG models are not adequately modelled by normally or log-normally distributed error. The important question for MEG error investigation would be how to represent the modelling uncertainty of the leadfield matrix. Of course there is always going to be some element of abstraction in the modelling of something as complex as the behaviour of magnetic fields in the human brain, but being able to account for this modelling uncertainty during the estimation process could lead to more accurate solutions.

The focus of our work on measurement-error was on its interaction with the sparse methods. In particular we looked at how the sparse methods were affected by measurement-error and how we could correct any effects. Consequently we have given little attention to the performance of the beamformer

under these circumstances. We briefly investigated the theoretical performance of the beamformer at the true location/orientation (see Appendix A), but this work could be extended to look at non-source locations or to include measurement-error in the beamformer. Also, as the beamformer can be viewed as a special case of a weighted least squares, it seems reasonable that we could modify our measurement-error methods so that they are suitable for beamformer correcting. In general, there is a lack of statistical theory around the beamformer. This is particularly important for multiple sources, where the impact of spatio-temporal factors on the beamforming performance is of interest. Zhang et al. (2014) (see also Zhang, 2015; Zhang and Liu, 2015) introduced a beamformer based on covariance thresholding and developed asymptotic theory for the relationship between beamformer performance and the spatial and temporal dimensions. Furthermore, Zhang (2015) proposed a threshold on the beamformer normalised power that was shown to have the sure screening property (Fan and Lv, 2008).

Finally, our implementation of the sparse methods in real MEG data raised the question of how to include depth/noise weighting into our methods. Since the sparse methods require linear programming schemes in order to arrive at a solution, the problem of introducing a weighting into the sparse estimates is not necessarily trivial. As we discussed in section 2.5, using a similar reasoning to the weighted beamformer/minimum norm leads to requiring an optimised weight matrix. An alternative approach is to only weight the penalty term, so that certain locations are more heavily penalised than others. An example of this is already present in the VESTAL method in the minimisation of a weighted ℓ_1 norm (Huang et al., 2006). This is a similar approach to the adaptive lasso

(Zou, 2006), which would probably be a good starting point for developing weighted sparse methods.

BIBLIOGRAPHY

Agogo, G., van der Voet, H., van't Veer, P., Ferrari, P., Muller, D., Sánchez-Cantalejo, E., Bamia, C., Braaten, T., Knüppel, S., Johansson, I., van Eeuwijk, F., and Boshuizen, H. (2016). A method for sensitivity analysis to assess the effects of measurement error in multiple exposure variables using external validation data. *BMC Medical Research Methodology*, 16:139–150.

Akalin Acar, Z. and Makeig, S. (2013). Effects of forward model errors on eeg source localization. *Brain Topography*, 26, 3:378–396.

Anderson, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society*, B, 38:1–36.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. John Wiley and Sons, Hoboken, third edition.

Belloni, A., Chernozhoukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98, 4:791–806.

Bickel, P. J., Ritov, Y. A., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, pages 1705–1732.

Biewen, E., Nolte (Lechner), S., and Rosemann, M. (2008). Multiplicative Measurement Error and the Simulation Extrapolation Method. Unpublished paper, `http://dx.doi.org/10.2139/ssrn.1131136`.

Brookes, M. J., Stevenson, C. M., Barnes, G. R., Hillebrand, A., Simpson, M. I., Francis, S. T., and Morris, P. G. (2007). Beamformer reconstruction of correlated sources using a modified source model. *Neuroimage*, 34 (4):1454–1465.

Brookes, M. J., Vrba, J., Robinson, S. E., Stevenson, C. M., Peters, A. M., Barnes, G. R., Hillebrand, A., and Morris, P. G. (2008). Optimising experimental design for MEG beamformer imaging. *Neuroimage*, 39 (4):1788–1802.

Brookes, M. J., Woolrich, M. W., and Price, D. (2014). An introduction to MEG connectivity measures. In Supek, S. and Aine, C., editors, *Magnetoencephalography: from signals to dynamic cortical networks*, pages 321–358. Springer.

Brookes, M. J., Zumer, J. M., Stevenson, C. M., Hale, J. R., Barnes, G. R., Vrba, J., and Morris, P. G. (2010). Investigating spatial specificity and data averaging in MEG. *Neuroimage*, 49 (1):525–538.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer, Heidelberg.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press, second edition.

Chatterjee, S. and Hadi, A. (2009). *Sensitivity analysis in linear regression*. John Wiley & Sons.

Cheyne, D., Bostan, A., Gaetz, W., and Pang, E. (2007). Event related beamforming: a robust method for presurgical functional mapping using MEG. *Clinical Neurophysiology*, 118:1691–1704.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association*, 89 (428):1314–1328.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., and Halgren, E. (2000). Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, 26 (1):55–67.

Dale, A. M. and Sereno, M. (1993). Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. *Journal of Cognitive Neuroscience*, 5:162–17.

David, O., Kilner, J. M., and Friston, K. J. (2006). Mechanisms of evoked and induced responses in MEG/EEG. *Neuroimage*, 31:1580–1591.

Dogandzic, A. and Nehorai, A. (2000). Estimating evoked dipole responses in unknown spatially correlated noise with EEG/MEG arrays. *IEEE Transactions on Signal Processing*, 48 (1):13–25.

Du, J. (2012). *Measurement error models in shape analysis*. PhD thesis, University of South Carolina.

Eckert, R. S., Carroll, R. J., and Wang, N. (1997). Transformations to additivity in measurement error models. *Biometrics*, 53:262–272.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32 (2):407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 456,:1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70 (5):849–911.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33 (1):1–22.

Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., Henson, R., Flandin, G., and Mattout, J. (2008). Multiple sparse priors for the M/EEG inverse problem. *Neuroimage*, 39:1104–1120.

Fuller, W. A. (1987). *Measurement error models*. John Wiley & Sons, New York.

Gonzalez-Moreno, A., Aurtenetxe, S., Lopez-Garcia, M.-E., del Pozo, F., Maestu, F., and Nevado, A. (2014). Signal-to-noise ratio of the MEG signal after preprocessing. *Journal of Neuroscience Methods*, 222:56–61.

Gramfort, A. and Kowalski, M. (2009). Improving M/EEG source localization with an inter-condition sparse prior. *IEEE Int Symp Biomed Imag*, pages 141–144.

Greene, W. H. (2008). *Econometric analysis*. Prentice Hall, New Jersey, sixth edition edition.

Greenland, S. (1996). Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*, 25, 6.

Hadamard, J. (1952). *Lectures on Cauchy's problem in linear partial differential equations*. Dover, New York.

Hamalainen, M. S. and Ilmoniemi, R. (1984). Interpreting measured magnetic fields of the brain: Estimates of current distributions. *Techical Report TKK-F-A559*.

Hamalainen, M. S., Lin, F., and Mosher, J. C. (2010). Anatomically and functionally constrained minimum-norm estimates. In Hansen, P., Kringelbach, M., and Salmelin, R., editors, *MEG: An introduction to methods*, pages 186–215. Oxford university press.

Hauk, O. (2004). Keep it simple: a case for using classical minimum norm estimation in the analysis of EEG and MEG data. *Neuroimage*, 21:1612–1621.

Herdman, A. T. and Cheyne, D. (2009). A practical guide for MEG and beamforming. In Handy, T., editor, *Brain signal analysis: advances in neuroelectric and neuromagnetic methods*, pages 99–140. MIT Press, Cambridge.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12:69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Huang, M. X., Dale, A. M., Song, T., Halgren, E., Harrington, D. L., Podgorny, I., Canive, J. M., Lewis, S., and Lee, R. R. (2006). Vector-based spatial-temporal minimum L1-norm solution for MEG. *Neuroimage*, 31 (3):1025–1037.

Huang, M. X., Huang, C. W., Robb, A., Angeles, A., Nichols, S. L., Baker, D. G., Song, T., Harrington, D. L., Theilmann, R. J., Srinivasan, R., and Heister, D. (2014). MEG source imaging method using fast L1 minimum-norm and its applications to signals with brain noise and human resting-state source amplitude images. *Neuroimage*, 84:585–604.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35, 1:73–101.

Huber, P. J. (1967). The behaviour of maximum likelihood estimates under non-standard conditions. *Proceedings of the 5th Berkeley Symposium*, 1:221–233.

Jia, J. and Yu, B. (2010). On model selection consistency of the elastic net when $p >> n$. *Statistica Sinica*, 20:595–611.

Karush, W. (1939). Minima of functions of several variables with inequalities as side constraints. M.sc. dissertation, Dept. of Mathematics, Univ. of Chicago, Chicago.

Kiebel, S. J., Daunizeau, J., Phillips, C., and Friston, K. J. (2008). Variational Bayesian inversion of the equivalent current dipole model in EEG/MEG. *Neuroimage*, 39 (2):728–741.

Kilner, J. M. and Friston, K. J. (2010). Topological inference for EEG and MEG. *The Annals of Applied Statistics*, pages 1272–1290.

Kipnis, V., Midthune, D., Freeman, L. S., Bingham, S., Day, N. E., Riboli, E., and Carroll, R. J. (2003). Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutrition*, 5:915–923.

Knight, K. and Fu, W. (2000). Asymptotics of lasso-type estimators. *The Annals of Statistics*, 28:1356–1378.

Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. In Neyman, J., editor, *Proceedings of 2nd Berkeley Symposium*, pages 481–492. University of California Press, Berkeley.

Lalancette, M., Quraan, M., and Cheyne, D. (2011). Evaluation of multiple-sphere head models for MEG source localization. *Physics in medicine and biology*, 56 (17):5621–5635.

Levine, D. (1985). The sensitivity of MLE to measurement error. *Journal of Econometrics*, 28:223–230.

Li, X., Zhao, T., Wang, L., Yuan, X., and Liu, H. (2014). Flare: Family of Lasso Regression, R package version 1.5.0., `https://cran.r-project.org/package=flare`.

Lin, F.-H., Witzel, T., Ahlfors, S. P., Stufflebeam, S. M., Belliveau, J. W., and Hamalainen, M. S. (2006). Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. *Neuroimage*, 31:160–171.

Litvak, V., Jha, A., Flandin, G., and Friston, K. (2013). Convolution models for induced electromagnetic responses. *Neuroimage*, 64:388–398.

Loh, P. L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40, 3:1637–1664.

Lopez, J. D., Penny, W. D., Espinosa, J. J., and Barnes, G. R. (2012). A general Bayesian treatment for MEG source reconstruction incorporating lead field uncertainty. *Neuroimage*, 60 (2-4):1194–1204.

Luckhoo, H. T., Brookes, M. J., and Woolrich, M. W. (2014). Multi-session statistics on beamformed MEG data. *Neuroimage*, 95:330–335.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34, 3:1436–1462.

Meyer, S. S., Bonaiuto, J., Lim, M., Rossiter, H., Waters, S., Bradbury, D., Bestmann, S., Brookes, M., Callaghan, M. F., Weiskopf, N., and Barnes, G. R. (2017). Flexible head-casts for high spatial precision MEG. *Journal of Neuroscience Methods*, 276:38–45.

Mosher, J. C., Leahy, R. M., and Lewis, P. S. (1999). EEG and MEG: forward solutions for inverse methods. *IEEE Transactions on Biomedical Engineering*, 46:245–259.

Mosher, J. C., Lewis, P. S., and Leahy, R. M. (1992). Multiple dipole modeling and localization from spatio-temporal MEG data. *IEEE Transactions on Biomedical Engineering*, 39:541–557.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32.

Nolte, S. (2007). The multiplicative simulation-extrapolation approach. Working Paper, `https://www.researchgate.net/publication/228820428_The_Multiplicative_Simulation-Extrapolation_Approach`.

Ou, W., Hamalainen, M. S., and Golland, P. (2009). A distributed spatio-temporal EEG/MEG inverse solver. *Neuroimage*, 44, 3:932–946.

Papanicolaou, A. C. (1998). *Fundamentals of functional brain imaging: A guide to the methods and their applications to psychology and behavioural neuroscience*. CRC Press.

Pascual-Marqui, R. D. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find. Exp. Clin. Pharmacol.*, 24(Suppl D.):5–12.

Pflug, G. C. (2002). Score function method. In *Encyclopedia of Statistical Sciences*. Kluwer Academic Publishers, Dordrecht.

Quenouille, M. (1956). Notes on Bias and estimation. *Biometrika*, 43:353–60.

Ramirez, R., Palmer, J., Makeig, S., Rao, B., and Wipf, D. (2007). Analysis of empirical Bayesian methods for neuroelectromagnetic source localization. *Advances in Neural Information Processing Systems*, pages 1505–1512.

Rao, C. R. (1973). *Linear statistical inference and its applications (2nd ed.)*. John Wiley & Sons, New York.

Robinson, S. E. (2004). Localization of event-related activity by SAM (erf). *Neurology and clinical neurophysiology: NCN*, pages 109–109.

Robinson, S. E. and Vrba, J. (1998). Functional neuroimaging by synthetic aperture magnetometry (SAM). In Yoshimoto, T., Kotani, M., Kuriki, S., Karibe, H., and Nakasato, N., editors, *Recent Advances in Biomagnetism*, pages 302–305. Sendai Tohoku University Press, Sendai.

Salmelin, R. and Parkkonen, L. (2010). Experimental design. In Hansen, P., Kringelbach, M., and Salmelin, R., editors, *MEG: An introduction to methods*, pages 75–82. Oxford university press.

Sekihara, K. and Nagarajan, S. S. (2015). *Electromagnetic brain imaging: A Bayesian perspective*. Springer.

Solin, A., Jylänki, P., Kauramäki, J., Heskes, T., van Gerven, M. A., and Särkkä, S. (2016). Regularizing solutions to the MEG inverse problem using space-time separable covariance functions. arXiv preprint arXiv:1604.04931.

Sørensen, Ø., Frigessi, A., and Thoresen, M. (2015). Measurement error in lasso: Impact and likelihood bias correction. *Statistica Sinica*, pages 809–829.

Sorrentino, A., Johansen, A. M., Aston, J. A. D., Nichols, T. E., and Kendall, W. S. (2013). Dynamic filtering of static dipoles in magnetoencephalography. *Annals of Applied Statistics*, 7:955–988.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, pages 703–716.

Stefanski, L. A. and Cook, J. R. (1995). Simulation extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90:1247–1256.

Stolk, A., Todorovic, A., Schoffelen, J.-M., and Oostenveld, R. (2013). Online and offline tools for head movement compensation in meg. *Neuroimage*, 68:39–48.

Tian, T. S., Huang, J. Z., Shen, H., and Li, Z. (2012). A two-way regularization method for MEG source reconstruction. *The Annals of Applied Statistics*, pages 1021–1046.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist.Soc.*, B 58:267–288.

Tikhonov, A. N. (1943). On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, 39 (5):195–198.

Uutela, K., Hamalainen, M., and Somersalo, E. (1999). Visualization of magnetoencephalographic data using minimum current estimates. *Neuroimage*, 10:173–180.

van de Geer, S. A. (2007). On non-asymptotic bounds for estimation in generalized linear models with highly correlated design. In *In Asymptotics: particles, processes and inverse problems*, pages 121–134. Institute of Mathematical Statistics.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Van Veen, B. D., van Drongelen, W., Yuchtman, M., and Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Trans. Biomed. Eng.*, 44 (9):867–880.

Vasiliu, D., Dey, T., and Dryden, I. L. (2014). Penalized Euclidean Distance Regression. arXiv preprint, `https://arxiv.org/abs/1405.4578`.

Vrba, J. (2002). Magnetoencephalography: the art of finding a needle in a haystack. *Physica*, c 368:1–9.

Wang, J. Z., Williamson, S. J., and Kaufman, L. (1993). Magnetic source imaging based on the minimum-norm least-squares inverse. *Brain Topography*, 5:365–371.

Waziri, A. E., Taggard, D. A., and Traynelis, V. C. (2005). Neurophysiology. In Moore, A.J., N. D., editor, *Neurosurgery: Principles and practice*, pages 3–23. Springer, London.

Wens, V., Marty, B., Mary, A., Bourguignon, M., Op de Beeck, M., Goldman, S., Van Bogaert, P., Peigneux, P., and De Tie, X. (2015). A geometric correction scheme for spatial leakage effects in MEG/EEG seed-based functional connectivity mapping. *Human Brain Mapping*, 36:4604–4621.

Willemsen, A., Frigo, C., and Boom, H. (1991). Lower extremity angle measurement with accelerometers-error and sensitivity analysis. *IEEE Transactions on Biomedical Engineering*, 38(12):1186–1193.

Yao, Z. and Eddy, W. F. (2014). A statistical approach to the inverse problem in magnetoencephalography. *The Annals of Applied Statistics*, 8 (2):1119–1144.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (1):49–67.

Zhang, J. (2015). On nonparametric feature filters in electromagnetic imaging. *Journal of Statistical Planning and Inference*, 164:39–53.

Zhang, J. and Liu, C. (2015). On linearly constrained minimum variance beamforming. *Journal of Machine Learning Research*, 16:2099–2145.

Zhang, J., Liu, C., and Green, G. (2014). Source localization with MEG data: A beamforming approach based on covariance thresholding. *Biometrics*, 70 (1):121–131.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101 (476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, B 67:301–320.

A

APPENDIX: BEAMFORMER DISTRIBUTIONAL THEORY

Since the beamformer method takes a very different approach to many, more frequently used statistical methods, it is of interest to investigate some theoretical performance of the method. Therefore, we now look to examine the distributional properties of the beamformer estimate. We will focus on the derivation of the approximate distribution at the true location/orientation, however the same reasoning can be followed to investigate non-source locations. We use the Taylor series expansion for the distribution approximation. If we consider the MEG model, we can write it as

$$\mathbf{d} = \mathbf{L}\mathbf{s} + \boldsymbol{\epsilon} = \boldsymbol{\mu} + \sigma\boldsymbol{\eta}$$

so that $\boldsymbol{\mu} = \mathbf{L}\mathbf{s}$, σ is the standard deviation of $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ is standard normally distributed. We want to find a Taylor approximation of

$$\hat{\mathbf{Q}}_\theta = \left(\mathbf{L}_\theta^T {\mathbf{C}_d}^{-1} \mathbf{L}_\theta\right)^{-1} \mathbf{L}_\theta^T {\mathbf{C}_d}^{-1} (\boldsymbol{\mu} + \sigma\boldsymbol{\eta}) = \mathbf{w}_\theta^T (\boldsymbol{\mu} + \sigma\boldsymbol{\eta})$$

where $\mathbf{C}_d = \frac{1}{T}\sum_{i=1}^{T} \mathbf{d}_i \mathbf{d}_i^T$. In the following expansion we will drop the subscript θ from vectors/matrices $\hat{\mathbf{Q}}, \mathbf{w}, \mathbf{L}$ for the sake of brevity. Nevertheless the expansion represents the approximate distribution for a given location/orientation θ.

Starting with the matrix $\mathbf{C}_d$,

$$\begin{aligned}\mathbf{C}_d &= \frac{1}{T}\sum_{i=1}^{T}\mathbf{d}_i\mathbf{d}_i^T = \frac{1}{T}\sum_{i=1}^{T}(\boldsymbol{\mu}+\sigma\boldsymbol{\eta}_i)(\boldsymbol{\mu}+\sigma\boldsymbol{\eta}_i)^T \\ &= \boldsymbol{\mu}\boldsymbol{\mu}^T + \sigma(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T) + \frac{\sigma^2}{T}\sum_{i=1}^{T}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T,\end{aligned}$$

and $\bar{\boldsymbol{\eta}} = \frac{1}{T}\sum_{i=1}^{T}\boldsymbol{\eta}_i$. Taking $\mathbf{B} = \boldsymbol{\mu}\boldsymbol{\mu}^T$, although in reality this will also be regularised, we now write $\mathbf{C}_d = \mathbf{B} + \sigma(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T) + \frac{\sigma^2}{T}\sum_{i=1}^{T}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T$. We can then take out a factor of $\mathbf{B}$ and subsequently use a power series expansion (up to terms of order σ^2) to expand the inverse of $\mathbf{C}_d$,

$$\begin{aligned}\mathbf{C}_d^{-1} &= \left[\mathbf{B}\left(\mathbf{I} + \sigma\mathbf{B}^{-1}(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T) + \frac{\sigma^2}{T}\mathbf{B}^{-1}\sum_{i=1}^{T}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T\right)\right]^{-1} \\ &= \mathbf{B}^{-1} - \sigma\mathbf{B}^{-1}(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T)\mathbf{B}^{-1} - \frac{\sigma^2}{T}\mathbf{B}^{-1}\left(\sum_{i=1}^{t}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T\right)\mathbf{B}^{-1} \\ &\quad + \sigma^2\mathbf{B}^{-1}(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T)\mathbf{B}^{-1}(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T)\mathbf{B}^{-1}.\end{aligned}$$

We then write this as $(\mathbf{B}^{-1} + \sigma\mathbf{D})$, where

$$\begin{aligned}\mathbf{D} = &-\mathbf{B}^{-1}(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T)\mathbf{B}^{-1} - \frac{\sigma}{T}\mathbf{B}^{-1}\left(\sum_{i=1}^{t}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T\right)\mathbf{B}^{-1} \\ &+ \sigma\mathbf{B}^{-1}(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T)\mathbf{B}^{-1}(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T)\mathbf{B}^{-1}.\end{aligned}$$

Then $\mathbf{L}^T\mathbf{C}_d^{-1}\mathbf{L} = (\mathbf{E} + \sigma\mathbf{L}^T\mathbf{D}\mathbf{L})$, where $\mathbf{E} = \mathbf{L}^T\mathbf{B}^{-1}\mathbf{L}$. Then using the same procedure as before, the inverse of $(\mathbf{L}^T\mathbf{C}_d^{-1}\mathbf{L})$ is expanded,

$$\begin{aligned}(\mathbf{L}^T\mathbf{C}_d^{-1}\mathbf{L})^{-1} &= (\mathbf{E} + \sigma\mathbf{L}^T\mathbf{D}\mathbf{L})^{-1} = \left[\mathbf{E}(\mathbf{I} + \sigma\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L})\right]^{-1} \\ &= \mathbf{E}^{-1} - \sigma\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1} + \sigma^2\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbf{w}^T &= (\mathbf{E}^{-1} - \sigma\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1} + \sigma^2\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}) \\ &\quad \times \mathbf{L}^T(\mathbf{B}^{-1} + \sigma\mathbf{D}) \\ &= \mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1} + \sigma\left[\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D} - \mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\right] \\ &\quad + \sigma^2\left[\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1} - \mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\right].\end{aligned}$$

The beamformer estimate for the i^{th} time point is then

$$\begin{aligned}
\hat{\mathbf{Q}}_i &= (\mathbf{L}^T\mathbf{C}_d^{-1}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{C}_d^{-1}\mathbf{d}_i \\
&= \Big(\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1} + \sigma\left[\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D} - \mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\right] \\
&\quad +\sigma^2\left[\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1} - \mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\right]\Big)(\mathbf{L}\mathbf{s}_i + \sigma\boldsymbol{\eta}_i) \\
&= \mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\mathbf{L}\mathbf{s}_i + \sigma\left[\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\boldsymbol{\eta}_i + \mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{s}_i\right. \\
&\quad \left.-\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\mathbf{L}\mathbf{s}_i\right] \\
&\quad + \sigma^2\left[\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\boldsymbol{\eta}_i + \mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\mathbf{L}\mathbf{s}_i\right. \\
&\quad \left.-\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\boldsymbol{\eta}_i - \mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{s}_i\right].
\end{aligned}$$

Recalling that $\mathbf{E} = \mathbf{L}^T\mathbf{B}^{-1}\mathbf{L}$, the first term will simplify, the second and third terms of σ and second and fourth terms of σ^2 will cancel. We are then left with,

$$\hat{\mathbf{Q}}_i = \mathbf{s}_i + \sigma\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\boldsymbol{\eta}_i + \sigma^2\left[\mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\boldsymbol{\eta}_i - \mathbf{E}^{-1}\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\boldsymbol{\eta}_i\right].$$

We now need to substitute $\mathbf{D}$ back into the expansion. However, since we are only working up to terms of order σ^2, the only term we need to be concerned with is $-\mathbf{B}^{-1}(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T)\mathbf{B}^{-1}$. Substituting this in, and using the fact that $\boldsymbol{\mu} = \mathbf{L}\mathbf{s}$, gives

$$\begin{aligned}
\{\hat{\mathbf{Q}}_i; O(\sigma^2)\} &= \mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T)\mathbf{B}^{-1}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\boldsymbol{\eta}_i \\
&\quad - \mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T)\mathbf{B}^{-1}\boldsymbol{\eta}_i \\
&= \mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}(\boldsymbol{\mu}\bar{\boldsymbol{\eta}}^T + \bar{\boldsymbol{\eta}}\boldsymbol{\mu}^T)\mathbf{B}^{-1}[\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1} - \mathbf{I}]\boldsymbol{\eta}_i \\
&= \mathbf{s}\bar{\boldsymbol{\eta}}^T\mathbf{B}^{-1}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\boldsymbol{\eta}_i - \mathbf{s}\bar{\boldsymbol{\eta}}^T\mathbf{B}^{-1}\boldsymbol{\eta}_i \\
&\quad + \mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\bar{\boldsymbol{\eta}}\mathbf{s}^T(\mathbf{L}^T\mathbf{B}^{-1}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1} - \mathbf{L}^T\mathbf{B}^{-1})\boldsymbol{\eta}_i \\
&= \mathbf{s}\bar{\boldsymbol{\eta}}^T\mathbf{B}^{-1}[\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1} - \mathbf{I}]\boldsymbol{\eta}_i,
\end{aligned}$$

using $\mathbf{E} = \mathbf{L}^T\mathbf{B}^{-1}\mathbf{L}$.

Therefore, the final expansion of the beamformer estimate up to σ^2 order terms is;

$$\hat{\mathbf{Q}}_i = \mathbf{s}_i + \sigma \mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\boldsymbol{\eta}_i + \sigma^2 \mathbf{s}\bar{\boldsymbol{\eta}}^T\mathbf{B}^{-1}[\mathbf{L}\mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1} - \mathbf{I}]\boldsymbol{\eta}_i,$$

where $\mathbf{B} = \mathbf{L}\mathbf{s}(\mathbf{L}\mathbf{s})^T$ and $\mathbf{E} = \mathbf{L}^T\mathbf{B}^{-1}\mathbf{L}$. Note, that these two matrices are the only two objects that need to be inverted through the process of the expansion. This therefore means that we require the matrix $\mathbf{B}$ to be invertible. In practice, this matrix will also incorporate any regularisation that is included in the covariance matrix.

The σ^2 term depends on the number of time points T through the sample mean $\bar{\boldsymbol{\eta}}$. Therefore for sufficiently large T, $\bar{\boldsymbol{\eta}} \to \mathbf{0}$ and hence the σ^2 term will approach zero. Furthermore, any terms of higher order of σ will be divided by powers of T and will also approach zero as our sample length increases.

The first order expansion then informs our approximate distribution,

$$\hat{\mathbf{Q}} \sim N(\mathbf{s}, \sigma^2 \mathbf{E}^{-1}\mathbf{L}^T\mathbf{B}^{-1}\mathbf{B}^{-1}\mathbf{L}\mathbf{E}^{-1}).$$