Supplementary Material to: Regularized Bidimensional Estimation of the Hazard Rate

Vivien Goepp[1], Jean-Christophe Thalabard[1], Grégory Nuel[2], and Olivier Bouaziz[1]

[1]*MAP5 (Department of mathematics, 45, rue des Saints-Pères, 75006 Paris)*
[2]*LPMA (Department of mathematics, 4, Place Jussieu, 75005 Paris)*

January 2019

S 1 Relation between the adaptive ridge and other L_1 and L_2 reweighted methods

As pointed out by a reviewer, Frommlet and Nuel (2016) provide no formal proof that the adaptive ridge approximates the L_q penalty for $q \in [0, 1)$. Other iterative methods, like Candès et al. (2008)'s L_1 reweighted scheme, have been shown to be an approximation of the L_q penalty using a Majorization-Minimization (MM, see Lange, 2004, Section 6) optimization scheme. This estimating procedure extends to the case $q = 0$, where it approximates the logarithmic penalty defined below. In this section, we show that the adaptive ridge minimizes the same function as the L_1 reweighted scheme. Both belong to the class of MM optimization algorithms and as such both are guaranteed to converge to a local minimum of the function to minimize.

We first prove that both methods solve the L_q penalty for $0 < q < 1$. We then show that they both extend to the case $q = 0$, where they are now approximations of the logarithm penalty instead.

S 1.1 MM Optimization for Solving L_q Penalties, $0 < q < 1$

Consider the problem of minimizing the likelihood penalized by the L_q norm:

$$\arg\min_{\boldsymbol{\beta}} \left\{ \ell_n(\boldsymbol{\beta}) + \frac{\kappa}{q} \|\boldsymbol{\beta}\|_q^q \right\} = \arg\min_{\boldsymbol{\beta}} \left\{ \ell_n(\boldsymbol{\beta}) + \frac{\kappa}{q} \sum_{j=1}^{p} |\beta_j|^q \right\}, \tag{1}$$

where $0 < q < 1$ and $\kappa > 0$ is the penalty constant, rescaled here by a factor $1/q$, and $\ell_n(\boldsymbol{\beta})$ is the function to minimize (in our case, the negative log-likelihood). This problem is difficult to solve because of the non-convexity of the L_q norm.

We will use MM optimization to derive a numerical scheme to solving Equation (1). MM Optimization makes use of a secondary function which majorizes the function to minimize (see

Hunter and Li, 2005). Since the majorization relation between functions is closed under sum, it suffices to focus in Equation (1) on the function $p(|\beta_j|) = |\beta_j|^q/q$ for $1 \leq j \leq p$ in order to construct an MM optimization scheme. We present two local approximations of $p(|\beta_j|)$ present in the literature, which give rise to two optimization schemes.

L_2 reweighted scheme Let $\beta_j^{(l)} \in \mathbb{R}$ be the current point of the numerical scheme. Using a local quadratic approximation (LQA, see Fan and Li, 2001; Hunter and Li, 2005), the function

$$q_{\text{LQA}}(\beta_j|\beta_j^{(l)}) = \frac{1}{2}\left|\beta_j^{(l)}\right|^{q-2}\beta_j^2 + \frac{2-q}{2q}\left|\beta_j^{(l)}\right|^q$$

majorizes $p(|\beta_j|)$ since we have $p(|\beta_j|) \leq q_{\text{LQA}}(\beta_j|\beta_j^{(l)})$ for every β_j with equality if and only if $\beta_j = \beta_j^{(l)}$. Define the current weights $w_j^{(l)} = |\beta_j^{(l)}|^{q-2}$. Noting that the second term of q_{LQA} does not depend on β_j, the MM optimization is given by the reweighted L_2 scheme:

$$\begin{aligned} \boldsymbol{\beta}^{(l)} &\leftarrow \arg\min_{\boldsymbol{\beta}} \left\{\ell_n(\boldsymbol{\beta}) + \frac{\kappa}{2}\sum_{j=1}^{d} w_j^{(l-1)}\beta_j^2\right\} \\ w_j^{(l)} &\leftarrow |\beta_j^{(l)}|^{q-2}, \end{aligned} \tag{2}$$

where (l) is the iteration index.

This scheme is the adaptive ridge procedure, where a small ε term is added to the reweighting step to bound the denominator away from zero (see discussion on this topic hereafter).

L_1 reweighted scheme Using a local linear approximation (LLA, see Zou and Li, 2008) the function,

$$q_{\text{LLA}}(\beta_j|\beta_j^{(l)}) = |\beta_j||\beta_j^{(l)}|^{q-1} + \frac{1-q}{q}|\beta_j^{(l)}|^q$$

majorizes $p(|\beta_j|)$. Defining now $w_j^{(l)} = |\beta_j^{(l)}|^{q-1}$, we obtain the following reweighted L_1 scheme:

$$\begin{aligned} \boldsymbol{\beta}^{(l)} &\leftarrow \arg\min_{\boldsymbol{\beta}} \left\{\ell_n(\boldsymbol{\beta}) + \kappa\sum_{j=1}^{d} w_j^{(l-1)}|\beta_j|\right\} \\ w_j^{(l)} &\leftarrow |\beta_j^{(l)}|^{q-1}. \end{aligned} \tag{3}$$

S 1.2 Extension to the case $q = 0$

Let us note that even though q has to be strictly positive in Equation (1), both numerical schemes (2) and (3) are still defined for $q = 0$. We now show that in the case $q = 0$, neither scheme solves the L_0 penalty: they correspond to a logarithmic penalty, which is a good approximation thereof (Candès et al., 2008, Section 2.3).

Let us first note that formally, the logarithmic penalty seems to be a good approximation to the L_0 norm since $\lim_{q\to 0}\sum_{j=1}^{p}|\beta_j|^q \to \|\boldsymbol{\beta}\|_0$ and $\lim_{q\to 0}(1/q)\sum_{j=1}^{p}(|\beta_j|^q - 1) = \sum_{j=1}^{p}\log(|\beta_j|)$.

In the context of sparse signal recovery, this is enough to prove that the logarithmic penalty yields the same estimate as the L_0 penalty (including the case with ε perturbation) (Wipf and Nagarajan, 2010, Section I). In the context of penalized likelihood, no such identity has been established. In this section, we make explicit the link between the logarithmic penalty and the L_0 penalty.

L_2 reweighted scheme We will start with the case of the L_2 reweighted scheme. Consider Problem (1) where the L_q penalty is replaced by the logarithmic penalty: $p(|\beta_j|) = \log(|\beta_j|)$. The LQA of this penalty around the current point $\beta_j^{(l)}$ is given by

$$\begin{aligned} q(\beta_j|\beta_j^{(l)}) &= p(|\beta_j^{(l)}|) + \left(\beta_j^2 - (\beta_j^{(l)})^2\right) \frac{p'(|\beta_j^{(l)}|)}{2|\beta_j^{(l)}|} \\ &= \frac{1}{2}\frac{\beta_j^2}{\left(\beta_j^{(l)}\right)^2} + \log(|\beta_j^{(l)}|) - \frac{1}{2} \end{aligned}$$

and the MM optimization scheme is obtained by iteratively minimizing $q(\beta_j|\beta_j^{(l)})$:

$$\boldsymbol{\beta}^{(l)} \leftarrow \arg\min_{\boldsymbol{\beta}} \left\{ \ell_n(\boldsymbol{\beta}) + \frac{\kappa}{2}\sum_{j=1}^{p} \frac{\beta_j^2}{\left(\beta_j^{(l)}\right)^2} \right\}, \tag{4}$$

which is the adaptive ridge with $q = 0$, with ε set to zero.

The case $\varepsilon > 0$, which is used in this paper, corresponds to the "square-log" penalty function $\beta_j \mapsto \log(1 + \beta_j^2/\varepsilon^2)$. More precisely, the pointwise convergence

$$\lim_{\varepsilon\to 0} \frac{1}{\log(1+1/\varepsilon^2)} \sum_{j=1}^{p} \log(1 + \beta_j^2/\varepsilon^2) = \|\boldsymbol{\beta}\|_0$$

justifies approximating Problem (1) when q is small by the problem

$$\arg\min_{\boldsymbol{\beta}} \left\{ \ell_n(\boldsymbol{\beta}) + \frac{\kappa}{\log(1+1/\varepsilon^2)} \sum_{j=1}^{p} \log(1 + \beta_j^2/\varepsilon^2) \right\}$$

when ε is small. MM optimization of the latter problem using LQA is obtained (after rescaling κ) by the update

$$\boldsymbol{\beta}^{(l)} \leftarrow \arg\min_{\boldsymbol{\beta}} \left\{ \ell_n(\boldsymbol{\beta}) + \frac{\kappa}{2}\sum_{j=1}^{p} \frac{\beta_j^2}{\left(\beta_j^{(l)}\right)^2 + \varepsilon^2} \right\},$$

which is the numerical scheme used in this paper. The theoretical properties of the ε-perturbed LQA are studied in Hunter and Li (2005) for a specific class of penalties but, to the best of the authors' knowledge, no theoretical properties have been established for the "square-log" penalty.

L_1 reweighted scheme The same reasoning applies to the LLA and shows (Zou and Li, 2008; Candès et al., 2008) that the L_1 reweighted scheme with $q = 0$ corresponds to the MM optimization of (1) with penalty function $p(|\beta_j|) = \log(|\beta_j|/|\varepsilon| + 1)$.

S 1.3 Related Works

Many works have made use of the reweighted L_1 and L_2 schemes derived above. We mention some related works of importance and finish with some remarks on the relative merits of the two methods.

These methods seem to first have been used in compressed sensing: Candès et al. (2008) studied the L_1 reweighted scheme with $q = 0$, while Daubechies et al. (2010) and Chartrand and Yin (2008) studied Algorithm (2) for $0 < q \leq 1$ and $q \in [0, 1)$ respectively. Johnson et al. (2012) studied Algorithm (3) with $q = 0$ in the context of linear regression. de Rooi and Eilers (2011), Rippe et al. (2012), and de Rooi et al. (2014) used Algorithm (2) with $q = 0$ in various applications, while Bach (2011, Section 5) and Mairal et al. (2014, Section 5.4) used the LQA to derive Algorithm (2) for the L_q penalty ($0 < q < 1$) and for more general norms. More recently, Frommlet and Nuel (2016) studied Algorithm (2) numerically for $q \in [0, 1)$, and specifically $q = 0$, under the name "adaptive ridge", which is the method used in this work. Dai et al. (2018) proved its consistency and oracle property in the setting of linear regression. Finally, Tardivel et al. (2018) have recently proven that, in the case of sparse signal recovery, the L_1 reweighted scheme with $q = 0$ is equivalent to minimizing the L_0 penalty problem.

Remark 1. The choice of q is independent from the choice between reweighted L_1 or L_2 schemes and it is not tackled here. Many papers cited in this section seem to favor choosing a small value of q.

Remark 2. Both reweighted L_1 and L_2 schemes have their advantages and drawbacks. The former is sparse at every step but each step requires solving a L_1 penalty. The latter is only asymptotically sparse and thus may require more iterations but it involves the simpler L_2 penalty, whose solution is explicit in the linear regression setting and simple to derive in other settings. To the best of our knowledge, there is no available implementation of the fused L_1 penalty for a general negative log-likelihood $\ell_n(\boldsymbol{\beta})$.

Remark 3. As in the present work, most works cited in this section use a modified weighting step for numerical stability: the denominator is bounded away from zero with an ε perturbation. While some offer rules of thumb to adaptatively decrease the value of ε as the algorithm converges, we have followed Frommlet and Nuel (2016)'s implementation and have set ε to a very small fixed value.

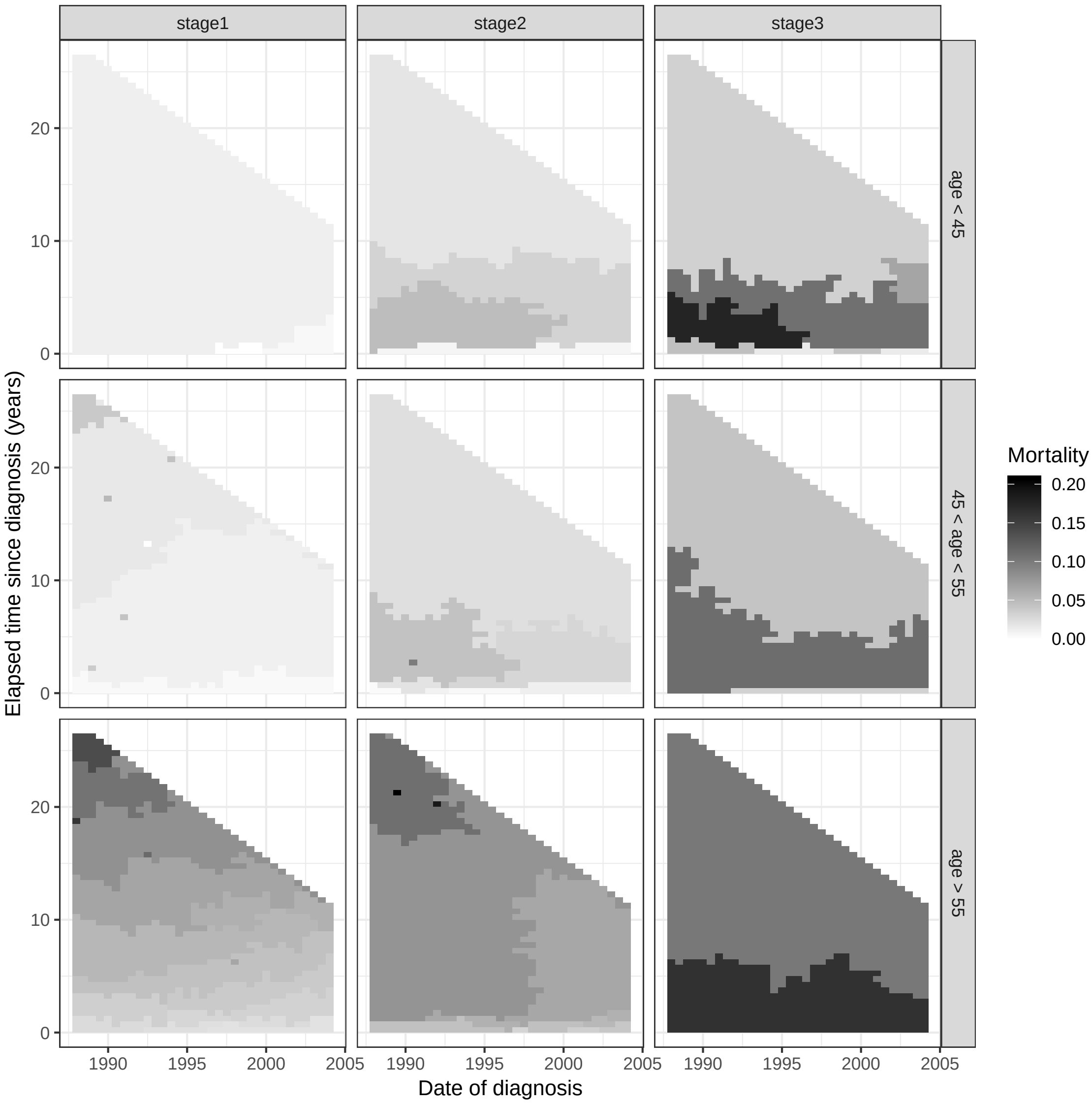

Figure 1: Estimated hazard of death since diagnosis of breast cancer for different cancer stages and for different ages at diagnosis. The estimate is obtained with the L_0 regularization. The upper right corner of every graph corresponds to the region where no data are available. All graphs share the same scale.

S 2 Application to Breast Cancer Mortality: Stratification with Respect to the Age at Diagnosis

The mortality of breast cancer is known to greatly vary on whether the cancer is pre or post-menopausal (Consensus, 1985). Consequently, a thorough analysis of the mortality from breast cancer would require to stratify with respect to the menopausal status at diagnosis. Since this covariate is not present in the data, we decided to stratify the sample with respect to the age of the patient at diagnosis, which is a proxy of menopausal status. Most women are known to have their menopause between 45 and 55 years old (Hill, 1996; Henderson et al., 2008; Gold, 2011), with 25th, 50th, 75th percentiles ranging from years 47-49, 50-51, 52-54, respectively, according to countries and surveys (Mishra et al., 2017). Consequently, based on the available information in SEER, for each cancer stage, the patients were divided into three classes of age at diagnosis: $(., 45]$, $(45, 55]$, and $(55, .)$ as a proxy for pre- menopausal, peri- menopausal and post- menopausal ages, respectively. The resulting estimated hazards are represented in Figure 1.

The stage 1 cancer patients younger than 45 and the stage 3 cancer patients older than 55 display the same mortality across all dates of diagnosis, i.e. with no cohort effect.

Moreover, the mortality of stage 1 cancer patients aged 45 and older at diagnosis has a slight cohort effect corresponding to a progressive decrease in the mortality across all survival times (Peto et al., 2000). This could suggest a trend of slow and steady improvement of the treatment of breast cancer in the United States over the period $1887 - 2005$.

Finally, we observe a clear decrease of the mortality for stage 2 cancers for all three age classes. This shift is located at the year 1995 for middle-aged patients and around the years $1997 - 1998$ for patients younger than 45 and older than 55. The same drop in mortality is observed for stage 3 cancers with patients younger than 45 at diagnosis, around year 1995. This could correspond to the introduction of improvements in the treatments of breast cancer in the United States (Consensus, 1985). Among the three main medical innovations, which can be considered in this period, the improvement of the surgical procedures for the loco- regional control of the disease and the assessment of the beneficial effect of hormone-receptor therapies could be reflected in the observed survival in stages 1-2, whereas the later emergence during this period of new classs of chemotherapeutic agents like taxoids (Rowinsky et al., 1992; Crown et al., 2004) or herceptin-based therapies targeted on new class of tumor markers (Pegram et al., 1998; Emens and Davidson, 2004) would be related with the changes in survival observed in stage 3. In the next section, we will use a stratified analysis to understand the effect of hormone-receptor therapies on the mortality shift in the mid-1990s.

S 3 Application to Breast Cancer Mortality: Stratification with Respect to the Estrogen Receptor Status

The cohort effect highlighted in the previous section could correspond to the introduction of Selective Estrogen Receptor Modulator (SERM) treatments and in particular the use of Tamoxifen as a treatment for breast cancer, showing improved survival in women with estrogen receptor positive tumor, initially in post- menopausal women (Fisher et al., 1989), later in both post - and pre-

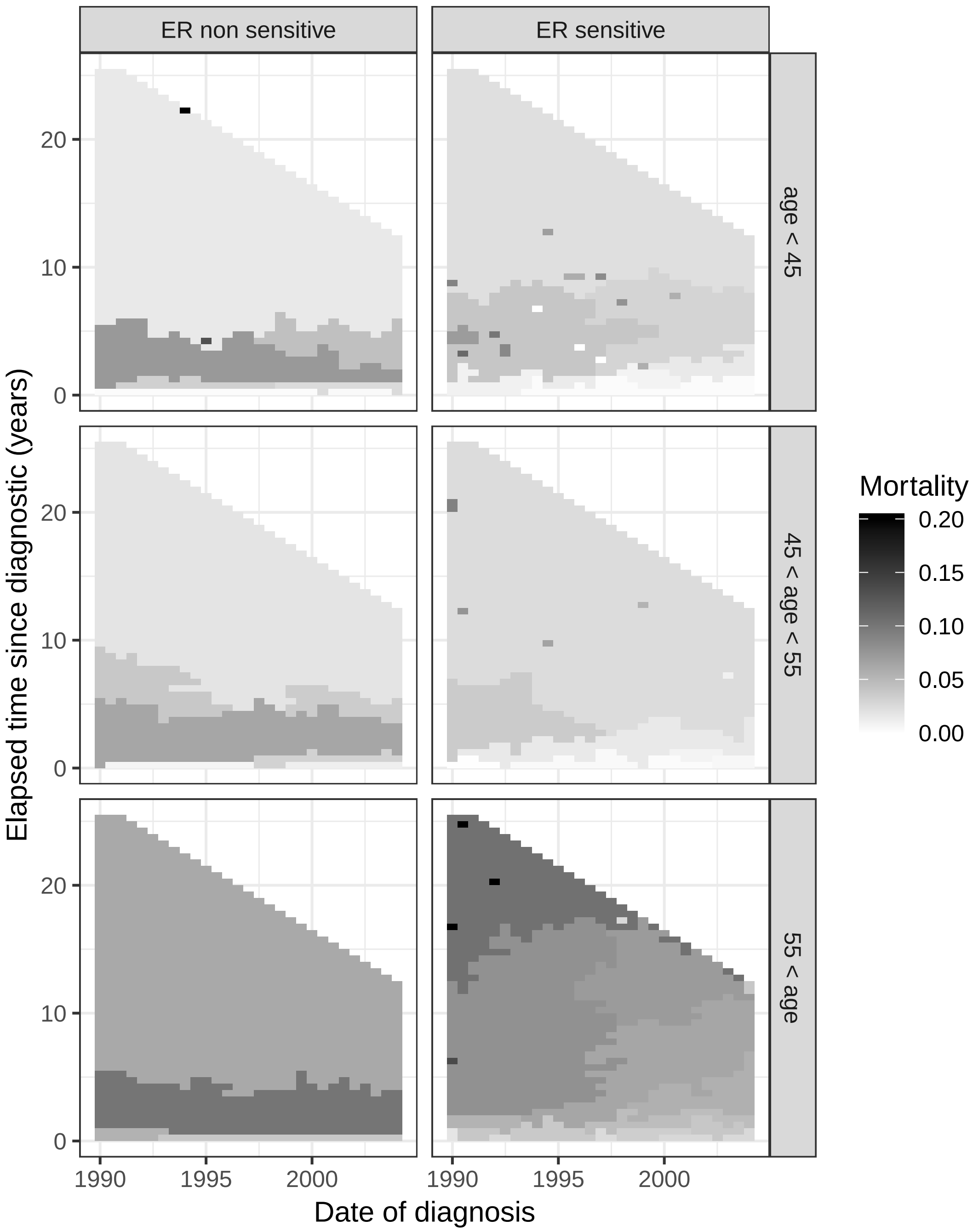

Figure 2: Estimated hazard of death since diagnosis of breast cancer for Stage 2 cancers. The estimation is carried separately for three classes of age at diagnosis: $(., 45]$, $(45, 55]$, and $(55, .)$ and for sensitive and non-sensitive estrogen receptor cancers. Inference is made with the L_0 regularization. All graphs share the same scale.

menopausal women (Early Breast Cancer Trialists' Collaborative Group, 1988; Fisher et al., 1998; Pritchard, 2005; Cochrane, 2008). Indeed, Tamoxifen was gradually used in the early years of 1990's (Gail et al., 1999; Harlan et al., 2002; Mariotto et al., 2006) to decrease the mortality of breast cancer patients. This treatment is only efficient on estrogen receptor-sensitive cancers. To validate our hypothesis, we conducted the estimation of mortality separately for patients with estrogen receptor sensitive and non-sensitive cancers. Since stage 2 cancers displayed a strong cohort effect across all ages at diagnosis, we only kept stage 2 cancers in this study. The estimated mortality is given in Figure 2. Note that the spikes in the mortality are an artifact of the segmentation procedure when the sample sizes tend to be too small in some regions of the age-cohort plane and are not to be taken into account in the interpretation of the mortality.

There is a clear difference in the evolution of mortality with respect to time at diagnosis between sensitive and non-sensitive estrogen cancers. For estrogen sensitive cases, the mortality displays the same sudden decrease around years $1997 - 1998$ as in Figure 1, across all age classes. In particular for individuals aged 55 or more at the time of diagnosis, the mortality has gradually decreased for estrogen sensitive patients, whereas it did not evolve with time for estrogen non-sensitive patients. On the other hand, the mortality for non-estrogen sensitive cancers displays almost no cohort effect for all ages at diagnosis (Knight et al., 1977).

The same analysis was run with stratification with respect to progesterone receptor status, with very similar morality estimates (results not shown here). Further analyses could be carried out to better understand the effect of the introduction of hormone-blocking therapies on mortality. However, the segmentation of the hazard rate, even with this simple stratified analysis, highlighted that the adoption of SERM therapies in the United States is a potential reason for the sharp decrease of mortality in the middle of the 1990s (Peto et al., 2000).

References

F. Bach. Optimization with Sparsity-Inducing Penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2011.

E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing Sparsity by Reweighted l1 Minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.

R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3869–3872, 2008.

Cochrane. Tamoxifen for early breast cancer. *The Cochrane database of systematic reviews*, (4): CD000486, 2008.

Consensus. Consensus conference. Adjuvant chemotherapy for breast cancer. *JAMA*, 254(24): 3461–3463, 1985.

J. Crown, M. O'Leary, and W.-S. Ooi. Docetaxel and paclitaxel in the treatment of breast cancer: A review of clinical experience. *The oncologist*, 9 Suppl 2:24–32, 2004.

L. Dai, K. Chen, Z. Sun, Z. Liu, and G. Li. Broken adaptive ridge regression and its asymptotic properties. *Journal of Multivariate Analysis*, 168:334–351, 2018.

I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1): 1–38, 2010.

J. de Rooi and P. Eilers. Deconvolution of pulse trains with the L0 penalty. *Analytica Chimica Acta*, 705(1-2):218–226, 2011.

J. J. de Rooi, C. Ruckebusch, and P. H. C. Eilers. Sparse Deconvolution in One and Two Dimensions: Applications in Endocrinology and Single-Molecule Fluorescence Imaging. *Analytical Chemistry*, 86(13):6291–6298, 2014.

Early Breast Cancer Trialists' Collaborative Group. Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. An overview of 61 randomized trials among 28,896 women. *The New England journal of medicine*, 319(26):1681–1692, 1988.

L. A. Emens and N. E. Davidson. Trastuzumab in breast cancer. *Oncology*, 18(9):1117–28; discussion 1131–2, 1137–8, 2004.

J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

B. Fisher, J. Costantino, C. Redmond, R. Poisson, D. Bowman, J. Couture, N. V. Dimitrov, N. Wolmark, D. L. Wickerham, and E. R. Fisher. A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. *The New England journal of medicine*, 320(8):479–484, 1989.

B. Fisher, J. P. Costantino, D. L. Wickerham, C. K. Redmond, M. Kavanah, W. M. Cronin, V. Vogel, A. Robidoux, N. Dimitrov, J. Atkins, M. Daly, S. Wieand, E. Tan-Chiu, L. Ford, and N. Wolmark. Tamoxifen for Prevention of Breast Cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *JNCI: Journal of the National Cancer Institute*, 90(18):1371–1388, 1998.

F. Frommlet and G. Nuel. An Adaptive Ridge Procedure for L0 Regularization. *PLoS ONE*, 11(2): e0148620, 2016.

M. H. Gail, J. P. Costantino, J. Bryant, R. Croyle, L. Freedman, K. Helzlsouer, and V. Vogel. Weighing the Risks and Benefits of Tamoxifen Treatment for Preventing Breast Cancer. *Journal of the National Cancer Institute*, 91(21):18, 1999.

E. B. Gold. The timing of the age at which natural menopause occurs. *Obstetrics and Gynecology Clinics of North America*, 38(3):425–440, 2011.

L. C. Harlan, J. Abrams, J. L. Warren, L. Clegg, J. Stevens, and R. Ballard-Barbash. Adjuvant therapy for breast cancer: Practice patterns of community physicians. *Journal of Clinical Oncology*, 20(7):1809–1817, 2002.

K. D. Henderson, L. Bernstein, B. Henderson, L. Kolonel, and M. C. Pike. Predictors of the timing of natural menopause in the Multiethnic Cohort Study. *American journal of epidemiology*, 167 (11):1287–1294, 2008.

K. Hill. The demography of menopause. *Maturitas*, 23(2):113–127, 1996.

D. R. Hunter and R. Li. Variable selection using MM algorithms. *The Annals of Statistics*, 33(4): 1617–1642, 2005.

B. A. Johnson, Q. Long, Y. Huang, K. Chansky, and M. Redman. Log-Penalized Least Squares, Iteratively Reweighted Lasso, and Variable Selection for Censored Lifetime Medical Cost. Technical Report Technical Report 2012-02, Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, 2012.

W. A. Knight, R. B. Livingston, E. J. Gregory, and W. L. McGuire. Estrogen receptor as an independent prognostic factor for early recurrence in breast cancer. *Cancer Research*, 37(12):4669–4671, 1977.

K. Lange. *Optimization*. Springer Texts in Statistics. Springer, New York, 2004.

J. Mairal, F. Bach, and J. Ponce. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

A. B. Mariotto, E. J. Feuer, L. C. Harlan, and J. Abrams. Dissemination of adjuvant multiagent chemotherapy and tamoxifen for breast cancer in the United States using estrogen receptor information: 1975-1999. *Journal of the National Cancer Institute. Monographs*, (36):7–15, 2006.

G. D. Mishra, N. Pandeya, A. J. Dobson, H.-F. Chung, D. Anderson, D. Kuh, S. Sandin, G. G. Giles, F. Bruinsma, K. Hayashi, J. S. Lee, H. Mizunuma, J. E. Cade, V. Burley, D. C. Greenwood, A. Goodman, M. K. Simonsen, H.-O. Adami, P. Demakakos, and E. Weiderpass. Early menarche, nulliparity and the risk for premature and early natural menopause. *Human Reproduction*, 32(3): 679–686, 2017.

M. D. Pegram, G. Pauletti, and D. J. Slamon. HER-2/neu as a predictive marker of response to breast cancer therapy. *Breast Cancer Research and Treatment*, 52(1-3):65–77, 1998.

R. Peto, J. Boreham, M. Clarke, C. Davies, and V. Beral. UK and USA breast cancer deaths down 25% in year 2000 at ages 20-69 years. *The Lancet*, 355(9217):1822, 2000.

K. Pritchard. Endocrinology and hormone therapy in breast cancer: Endocrine therapy in premenopausal women. *Breast Cancer Research*, 7(2):70–76, 2005.

R. C. A. Rippe, J. J. Meulman, and P. H. C. Eilers. Visualization of Genomic Changes by Segmented Smoothing Using an L0 Penalty. *PLoS ONE*, 7(6):e38230, 2012.

E. K. Rowinsky, N. Onetto, R. M. Canetta, and S. G. Arbuck. Taxol: The first of the taxanes, an important new class of antitumor agents. *Seminars in Oncology*, 19(6):646–662, 1992.

P. J. C. Tardivel, R. Servien, and D. Concordet. Sparsest representations and approximations of an underdetermined linear system. *Inverse Problems*, 34(5):055002, 2018.

D. Wipf and S. Nagarajan. Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):317–329, 2010.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008.