Noname manuscript No.
(will be inserted by the editor)

Data-adaptive Shrinkage via the Hyperpenalized EM Algorithm

Philip S. Boonstra · Jeremy M. G. Taylor · Bhramar Mukherjee

the date of receipt and acceptance should be inserted later

Abstract We propose an extension of the expectation-maximization (EM) algorithm, called the hyperpenalized EM (HEM) algorithm, that maximizes a penalized log-likelihood, for which some data are missing or unavailable, using a data-adaptive estimate of the penalty parameter. This is potentially useful in applications for which the analyst is unable or unwilling to choose a single value of a penalty parameter but instead can posit a plausible range of values. The HEM algorithm is conceptually straightforward and also very effective, and we demonstrate its utility in the analysis of a genomic data set. Gene expression measurements and clinical covariates were used to predict survival time. However, many survival times are censored, and some observations only

Department of Biostatistics, University of Michigan, 1415 Washington Hts., Ann Arbor, MI, USA.

E-mail: philb@umich.edu

Tel. +1 (734) 615-1580

contain expression measurements derived from a different assay, which together constitute a difficult missing data problem. It is desired to shrink the genomic contribution in a data-adaptive way. The HEM algorithm successfully handles both the missing data and shrinkage aspects of the problem.

Keywords EM algorithm · hyperparameter · hyperpenalty · missing data · penalized likelihood · prediction

1 Introduction

Since its formal description by Dempster et al. (1977), the expectation-maximization (EM) algorithm has played a fundamental role in missing-data problems. This is primarily due to its desirable theoretical properties and extensive applicability, established both by Dempster et al. and others, including Wu (1983), Meng and Rubin (1993), and Van Dyk (2000), among others. The EM algorithm is effective for fitting a statistical model, the likelihood of which is difficult or even intractable to maximize but would be made feasible given additional missing data or latent variables. Many papers have extended the EM algorithm to further its utility; among these is the penalized EM (PEM) algorithm (Green, 1990), which allows for the maximization of a difficult-to-calculate *penalized* likelihood. Although not of primary interest to Green, an important component of the penalized likelihood is the choice of penalty parameter. In this paper, we extend the EM and PEM algorithms to simultaneously allow for the selection of this penalty parameter.

We introduce notation to formalize our objective and provide background to the problem. Let $\boldsymbol{U}^{\text{obs}}$ denote observed data and $\boldsymbol{\theta}$ the set of model parameters. Using the convention "$[\cdot]$" and "$[\cdot|\cdot]$" to represent marginal and conditional density functions, respectively, the EM algorithm indirectly maximizes a difficult or intractable observed-data log-likelihood, $\ell_O(\boldsymbol{\theta}) = \ln[\boldsymbol{U}^{\text{obs}}|\boldsymbol{\theta}]$, with respect to $\boldsymbol{\theta}$. It does so by introducing missing data $\boldsymbol{U}^{\text{mis}}$ in such a way that the *complete*-data log-likelihood, $\ell_C(\boldsymbol{\theta}) = \ln[\boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}}|\boldsymbol{\theta}]$, is easier to calculate. The quantity $\boldsymbol{U}^{\text{mis}}$ may be genuinely missing, such as the unobserved value of a censored outcome, or latently missing, such as cluster membership in a posited mixture model.

Mechanistically, $\ell_O(\boldsymbol{\theta})$ is indirectly maximized through successive iterations of E- and M-steps:

EM algorithm (Dempster et al., 1977)

$$\begin{aligned} &\textbf{E-step: } Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}(\ell_C(\boldsymbol{\theta})|\boldsymbol{\theta}^{(t)}, \boldsymbol{U}^{\text{obs}}) \\ &\textbf{M-step: } \boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \end{aligned} \tag{1}$$

Here and throughout the paper, $\boldsymbol{\theta}^{(t)}$ is the value of $\boldsymbol{\theta}$ at iteration t. The crucial result from Dempster et al. (1977) is that each iteration does not decrease the observed log-likelihood: $\ell_O(\boldsymbol{\theta}^{(t+1)}) \geq \ell_O(\boldsymbol{\theta}^{(t)})$. The inequality is strict if $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) > Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$, that is, if the expected complete log-likelihood is increased between iterations. If $\ell_O(\boldsymbol{\theta})$ is unimodal with one stationary point, then, under first-order differentiability assumptions on $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, the sequence $\{\boldsymbol{\theta}^{(t)}\}$ from an EM algorithm converges to $\operatorname{argmax}_{\boldsymbol{\theta}} \ell_O(\boldsymbol{\theta})$, the maximum like-

lihood estimate of $\boldsymbol{\theta}$ (Dempster et al., 1977; Wu, 1983). In practice, multiple starting points for the EM algorithm are recommended to check for multimodality.

The PEM algorithm leaves the E-step unchanged and modifies the M-step:

PEM algorithm (Green, 1990)

E-step: $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}(\ell_C(\boldsymbol{\theta})|\boldsymbol{\theta}^{(t)}, \boldsymbol{U}^{\mathrm{obs}})$

PM-step: $\boldsymbol{\theta}^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\theta}} \{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - p_{\boldsymbol{\eta}}(\boldsymbol{\theta})\}$

The function $p_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ is a convex penalty function indexed by $\boldsymbol{\eta}$, which represents the penalty parameter(s). The sequence $\{\boldsymbol{\theta}^{(t)}\}$ now converges to $\mathrm{argmax}_{\boldsymbol{\theta}} \{l_O(\boldsymbol{\theta}) - p_{\boldsymbol{\eta}}(\boldsymbol{\theta})\}$. Green considers several applications of the PEM algorithm for which values of $\boldsymbol{\eta}$ are pre-selected. In many non-missing-data-scenarios, however, a penalty function is introduced to achieve shrinkage (ridge regression, Hoerl and Kennard, 1970), variable selection (lasso, Tibshirani, 1996) or simultaneous shrinkage and variable selection (elastic net, Zou and Hastie, 2005). In these cases, $\boldsymbol{\eta}$ is a tuning parameter, the choice of which is typically data-driven, with cross validation or generalized cross-validation being standard approaches (e.g. Hastie et al., 2009). It is not clear however, how such strategies would translate in missing-data scenarios. That is, cross-validation repeatedly partitions the data into independent training and "left-out" subsets, using the latter to choose the tuning parameter without overfitting. An imputation strategy to address any missing data would need to be separately applied to each subset, so as to maintain independence, but this could cause a potentially large loss of

efficiency. The objective of this paper is addressing the challenge of choosing $\boldsymbol{\eta}$ within the EM context.

In the original EM paper, Dempster et al. (1977) indirectly take up the problem of choosing $\boldsymbol{\eta}$ in a demonstrative application of the algorithm to a hierarchical model. They move the model parameters $\boldsymbol{\theta}$ to the E-step, i.e. treat $\boldsymbol{\theta}$ as "missing" data, and use the M-step to estimate hyperparameters, $\boldsymbol{\eta}$. Thus, the observed likelihood is considered a function of $\boldsymbol{\eta}$ alone, as in a variance components model, and is maximized with respect to $\boldsymbol{\eta}$. To be clear, the only "missing data" in that specific example was $\boldsymbol{\theta}$. If, in addition to $\boldsymbol{\theta}$, there is also actual missing data $\boldsymbol{U}^{\text{mis}}$, the E-step in (1) may become considerably more difficult to calculate, requiring an expectation with respect to the joint density of $\{\boldsymbol{U}^{\text{mis}}, \boldsymbol{\theta}\}$ given $\{\boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}\}$. Also, this example does not estimate a posterior mode of $\boldsymbol{\theta}$ but instead treats $\boldsymbol{\eta}$ as the main parameter of interest.

The ability to move $\boldsymbol{\theta}$ from the M- to the E-step blurs the distinction between "missing data" and "unknown parameter" and is related to a fully Bayesian data augmentation approach (Tanner and Wong, 1987) or partially Bayesian approaches that incorporate Monte Carlo methods to numerically approximate the E-step (Wei and Tanner, 1990; Casella, 2001). From a Bayesian perspective, the penalty function $p_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ corresponds to a prior on $\boldsymbol{\theta}$ indexed by hyperparameters $\boldsymbol{\eta}$, and a Bayesian treatment of the problem iteratively samples $\boldsymbol{U}^{\text{mis}}$ and $\boldsymbol{\theta}$ from their respective conditional distributions. Such an algorithm will eventually yield a draw from the posterior, $[\boldsymbol{\theta}, \boldsymbol{U}^{\text{mis}} | \boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}]$. Extending the Bayesian approach when $\boldsymbol{\eta}$ is unknown, as in our situation, Gelfand and

Smith (1990) showed how $\boldsymbol{\eta}$ may also be sampled from *its* conditional distribution, i.e. $[\boldsymbol{\eta}|\boldsymbol{\theta}]$. Boonstra et al. (2014a) provide an in-depth comparison of these Bayesian approaches. In contrast, the PEM algorithm seeks a posterior mode: $\text{argmax}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}] = \text{argmax}_{\boldsymbol{\theta}} \{l_O(\boldsymbol{\theta}) - p_{\boldsymbol{\eta}}(\boldsymbol{\theta})\}$.

It is within this framework of the Bayesian approach (hierarchical modeling) intersecting with the EM algorithm (point estimation with missing data) that our extension fits. The *hyperpenalized* EM (HEM) algorithm allows $\boldsymbol{\eta}$ to be unknown, in contrast with the PEM algorithm, and gives support to a range of values by way of a so-called hyperpenalty (Boonstra et al., 2014b), $h(\boldsymbol{\eta})$, which penalizes the penalty parameter itself. Boonstra et al. introduced the use of a hyperpenalty when fitting a ridge regression with a small sample size, for which cross-validation is susceptible to overfitting, and showed by way of simulation study and a data analysis that the hyperpenalty approach was frequently superior to nine other existing methods, including five-fold cross-validation and generalized cross-validation, in terms of prediction. Here, we extend the concept to a missing data context within a more general penalized likelihood framework beyond ridge regression. In the HEM algorithm, both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are updated sequentially by maximization steps. One iteration proceeds

as follows:

HEM algorithm

E-step $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}(\ell_C|\boldsymbol{\theta}^{(t)}, \boldsymbol{U}^{\mathrm{obs}})$.

M-step $\boldsymbol{\theta}^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\theta}} \{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - p_{\boldsymbol{\eta}^{(t)}}(\boldsymbol{\theta})\}$. (2)

H-step $\boldsymbol{\eta}^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\eta}} \{-p_{\boldsymbol{\eta}}(\boldsymbol{\theta}^{(t+1)}) - h(\boldsymbol{\eta})\}$. (3)

Like the PEM algorithm, the HEM algorithm does not comprise a fully Bayesian analysis, because it gives point estimates, namely posterior modes, rather than an estimate of the entire distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Put differently, the HEM algorithm is to a fully Bayesian analysis with a hyperprior on $\boldsymbol{\eta}$ as the PEM algorithm is to a fully Bayesian analysis with a fixed value of $\boldsymbol{\eta}$. The hyperpenalty serves the role of the hyperprior in a Bayesian analysis. This presents a tradeoff: the HEM algorithm provides only a single point estimate of $\boldsymbol{\theta}$, but a posterior mode can often be numerically calculated quickly relative to estimating the entire posterior distribution, particularly for a complicated, non-conjugate hyperpenalty and/or when there is a large fraction of missing data.

The remainder of this paper proceeds as follows. First, we establish some basic properties of the HEM algorithm and provide guidance on appropriate choices of the hyperpenalty $h(\boldsymbol{\eta})$ when the penalty $p_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ belongs to the exponential family of distributions, which, as we will see, includes the cases of ridge regression, the lasso, and the elastic net (Section 2). We then consider two applications of the HEM algorithm. The first, in Section 3, is demonstrative in

nature, to compare the HEM to the (P)EM algorithms. The second, in Section 4, comes from a recent analysis with a large amount of missing data that also required adaptive shrinkage of regression coefficients (Chen et al., 2011), for which the (P)EM algorithms are ill-equipped. We will close with a discussion in Section 5.

2 Hyperpenalty

The EM algorithm is based on the likelihood decomposition $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \ell_O(\boldsymbol{\theta}) + H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, with $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}(\ln[\boldsymbol{U}^{\mathrm{mis}}|\boldsymbol{U}^{\mathrm{obs}}, \boldsymbol{\theta}]|\boldsymbol{U}^{\mathrm{obs}}, \boldsymbol{\theta}^{(t)})$ and the dependence on $\boldsymbol{U}^{\mathrm{obs}}$ of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ and $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ being notationally suppressed. Jensen's inequality gives that $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \leq H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ for any $\boldsymbol{\theta}$, and so increasing Q at each iteration also increases ℓ_O. We can use this relation to establish an analogous result for the HEM algorithm:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - p_{\boldsymbol{\eta}}(\boldsymbol{\theta}) - h(\boldsymbol{\eta}) = \ell_O(\boldsymbol{\theta}) + H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - p_{\boldsymbol{\eta}}(\boldsymbol{\theta}) - h(\boldsymbol{\eta}).$$

Thus, increasing the *expected* "hyperpenalized log-likelihood" (HLL), namely $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - p_{\boldsymbol{\eta}}(\boldsymbol{\theta}) - h(\boldsymbol{\eta})$, also increases the *observed* HLL, $\ell_O(\boldsymbol{\theta}) - p_{\boldsymbol{\eta}}(\boldsymbol{\theta}) - h(\boldsymbol{\eta})$. The HEM algorithm maintains this desirable feature of the EM and PEM algorithms. We next discuss some minimal conditions required for a hyperpenalty.

It is reasonable to require that any particular choice of $h(\boldsymbol{\eta})$ not give rise to an infinite expected HLL. For example, in ridge regression (Hoerl and Kennard, 1970), the penalty corresponds to a normal distribution, and, up to an additive

constant, $p_{\boldsymbol{\eta}}(\boldsymbol{\theta}) = \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}/2 - k\ln(\lambda)/2$, where $\boldsymbol{\eta} = \{\lambda\}$, with λ a 1-dimensional tuning parameter, and $\boldsymbol{\theta} = \{\boldsymbol{\beta}\}$, with $\boldsymbol{\beta}$ the k-dimensional vector of regression coefficients (we assume here the error variance is known). In this case, $h(\boldsymbol{\eta}) \equiv h(\lambda)$. At iteration t, the HEM algorithm sequentially determines values of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ that maximize the expected HLL, which in this case is

$$Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) + k\ln(\lambda)/2 - \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}/2 - h(\lambda). \tag{4}$$

The $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)})$ expression is left in general terms; the exact nature of the missing data is not relevant to this example, except for assuming it is finite for any finite value of $\boldsymbol{\beta}$. In the extreme case of an uninformative hyperpenalty, say $h(\lambda) = C$, where C is a constant, the expression in (4) can be made arbitrarily large with λ by setting $\boldsymbol{\beta} = \mathbf{0}_k$. To prevent this, $h(\lambda)$ must approach ∞ sufficiently fast with λ. The following result establishes a sufficient condition to ensure a finite expected HLL under a special class of penalty functions; the proof is given in Web Appendix A. We assume here that $\boldsymbol{\eta} = \{\lambda, \alpha\}$, where λ is the unknown tuning parameter and α is fixed or non-existent. We will show that this class includes several typical choices of $p_{\boldsymbol{\eta}}(\boldsymbol{\theta})$.

Result Suppose that the penalty function belongs to a special class of the exponential family of distributions:

$$\exp\{-p_{\boldsymbol{\eta}}(\boldsymbol{\theta})\} \equiv \exp\{-p_{\lambda,\alpha}(\boldsymbol{\theta})\} = \exp\{-\lambda f(\boldsymbol{\theta}, \alpha) + g(\lambda, \alpha)\}, \tag{5}$$

with $\boldsymbol{\theta}$ as the random variable and $\boldsymbol{\eta} = \{\lambda, \alpha\}$ the set of parameters, where α is known and $\lambda \geq 0$ is the unknown canonical parameter. Suppose also that $f(\boldsymbol{\theta}, \alpha) \geq 0$ for all $\boldsymbol{\theta}$ and α, with equality if and only if $\boldsymbol{\theta} = \mathbf{0}_k$. Thus, larger

values of λ induce a greater penalty on $\boldsymbol{\theta}$. Let $h(\lambda, \alpha)$ be a continuous function such that $\lim_{\lambda \to 0}\{-h(\lambda, \alpha)\} < \infty$ and $\lim_{\lambda \to \infty}\{g(\lambda, \alpha) - h(\lambda, \alpha)\} < \infty$. Then, for an arbitrary value of $\boldsymbol{\theta}$, $\max_{\lambda}\{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - p_{\lambda,\alpha}(\boldsymbol{\theta}) - h(\lambda, \alpha)\} < \infty$.

Remark 1 The requirement that $\lim_{\lambda \to 0}\{-h(\lambda, \alpha)\} < \infty$ is trivial, because a violation to this assumption implies that λ must be zero.

Remark 2 A consequence of this result is that even a hyperpenalty that corresponds to a proper hyperprior on λ, i.e. $\int_{\lambda} \exp\{-h(\lambda, \alpha)\}\mathrm{d}\lambda < \infty$, does not guarantee that the HEM algorithm won't diverge. Rather, $h(\lambda, \alpha)$ must be decreasing with λ at a rate that depends on the specific choice of penalty function $p_{\lambda,\alpha}(\boldsymbol{\theta})$.

Remark 3 When λ and α are both fixed, meaning there is no need for the HEM algorithm, the normalizing value $g(\lambda, \alpha)$ is constant and may be ignored, reducing Equation (5) to the more familiar $\exp\{-p_{\lambda,\alpha}(\boldsymbol{\theta})\} = \exp\{-\lambda f(\boldsymbol{\theta}, \alpha)\}$. When λ is unknown, as in the setting of this paper, $g(\lambda, \alpha)$ must be included in the expression for $\exp\{-p_{\lambda,\alpha}(\boldsymbol{\theta})\}$ in order to remain (proportional to) a probability density function.

In the case of penalized linear regression, this result encompasses several common choices of penalty, including ridge regression, which was already discussed: $f(\boldsymbol{\theta}, \alpha) = \boldsymbol{\beta}^{\top}\boldsymbol{\beta}/2$ and $g(\lambda, \alpha) = k\ln(\lambda/[2\pi])/2$. This result also includes the lasso (Tibshirani, 1996), for which the penalty coincides with the Laplace distribution: $f(\boldsymbol{\theta}, \alpha) = \|\boldsymbol{\beta}\|$ and $g(\lambda, \alpha) = k\ln(\lambda/2)$. In Web Appendix B, we derive the normalizing constant for the distribution that conforms to the

elastic-net penalty (Zou and Hastie, 2005), which is a convex combination of the ridge and lasso penalties controlled by a weighting parameter α. The above result applies to this penalty also:

$$-p_{\lambda,\alpha}(\boldsymbol{\theta}) = -\lambda\left(\alpha\|\boldsymbol{\beta}\| + (1-\alpha)\boldsymbol{\beta}^\top\boldsymbol{\beta}/2\right) \\ - \lambda k\alpha^2/(2[1-\alpha]) + k\ln(\lambda)/2 - k\ln\Phi\left(-\alpha\sqrt{\lambda}/\sqrt{1-\alpha}\right) - k\ln(8\pi/[1-\alpha])/2. \tag{6}$$

The function $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. Thus, for the elastic net $f(\boldsymbol{\theta},\alpha) = \alpha\|\boldsymbol{\beta}\| + (1-\alpha)\boldsymbol{\beta}^\top\boldsymbol{\beta}/2$, and the second row of (6) comprises $g(\lambda,\alpha)$. A hyperpenalty could be used to simultaneously select both α and λ, but the result above would need to be made more general to ensure a finite HLL.

Using these results, we next apply the HEM algorithm to two examples.

3 Example 1: Multinomial Distribution

This first example of the HEM algorithm continues the demonstrative application from Dempster et al. (1977), which was extended in Green (1990). The unobserved complete data are $\boldsymbol{x}^\top = \{x_1, x_2, x_3, x_4, x_5\}$, assumed to be generated from a multinomial distribution of 197 independent trials, $M\{197; 1/2, \zeta/4, (1/4)(1-\zeta), (1/4)(1-\zeta), \zeta/4\}$, with the unknown parameter $\zeta \in (0,1)$. However, we only observe $\boldsymbol{y}^\top = \{x_1 + x_2, x_3, x_4, x_5\} = \{125, 18, 20, 34\}$, coming from $M\{197; 1/2+\zeta/4, (1/4)(1-\zeta), (1/4)(1-\zeta), \zeta/4\}$. Thus, $\boldsymbol{U}_{\text{obs}} = \boldsymbol{y}$, $\boldsymbol{U}_{\text{mis}} = x_2$, and the complete data log-likelihood is $\ell_C(\zeta) = (x_2+x_5)\ln\zeta + (x_3+x_4)\ln(1-$

ζ). The distribution of x_2 given $x_1 + x_2 = 125$ is Binomial: $B\{125; \zeta/(2+\zeta)\}$, and the E-step simplifies to calculating $\mathrm{E}(x_2|\boldsymbol{y}, \zeta^{(t)}) = 125 \cdot \zeta^{(t)}/(2 + \zeta^{(t)})$. Green (1990) penalized the log-likelihood with $p_\lambda(\zeta) = 10(\zeta - 0.5)^2$, the negative log-density of a Normal prior with mean 0.5 and precision $\lambda = 20$, truncated to the $(0, 1)$ interval. We compare the resulting analyses from the original EM algorithm, the PEM algorithm, and an implementation of the HEM algorithm that allows for the precision λ to be unknown, described as follows. The M-step from the original EM algorithm solves a linear equation of ζ:

$$\textbf{EM} \quad \frac{\partial}{\partial\zeta} Q(\zeta|\zeta^{(t)}) = \frac{\mathrm{E}(x_2|\boldsymbol{y}, \zeta^{(t)}) + x_5}{\zeta} + \frac{x_3 + x_4}{1 - \zeta} = 0.$$

The penalized M-step from Green solves the following for $\zeta \in (0, 1)$:

$$\textbf{PEM} \quad \frac{\partial}{\partial\zeta} \left(Q(\zeta|\zeta^{(t)}) - p_{\lambda=20}(\zeta) \right) = \frac{\mathrm{E}(x_2|\boldsymbol{y}, \zeta^{(t)}) + x_5}{\zeta} + \frac{x_3 + x_4}{1 - \zeta} - 20(\zeta - 0.5) = 0.$$

In the HEM algorithm implementation, we leave the precision of the prior, formerly $\lambda = 20$, unspecified. We give support to a range of values using the hyperpenalty $h(\lambda) = -(a-1)\ln(\lambda) + b\lambda$. This choice corresponds to a Gamma random variable with moments $\mathrm{E}(\lambda) = a/b$ and $\mathrm{Var}[\lambda] = a/b^2$ and satisfies the requirements of the result in Section 2 for any $a, b > 0$. Using $a = 4$ and $b = 0.2$ gives $\mathrm{E}(\lambda) = 20$ and $\mathrm{Var}[\lambda] = 100$ and allows for more uncertainty about an appropriate choice of λ relative to fixing $\lambda = 20$. The updates for ζ and λ are, respectively:

$$\textbf{HEM} - \textbf{M-step} \quad \frac{\mathrm{E}(x_2|\boldsymbol{y}, \zeta^{(t)}) + x_5}{\zeta} + \frac{x_3 + x_4}{1 - \zeta} - \lambda^{(t)}(\zeta - 0.5) = 0,$$

$$\textbf{HEM} - \textbf{H-step} \quad 1/(2\lambda) - \frac{1}{2}(\zeta^{(t+1)} - 0.5)^2 - \frac{\phi(\sqrt{\lambda}/2)\lambda^{-1/2}}{4\Phi(\sqrt{\lambda}/2) - 2} + 3/\lambda - 1/5 = 0.$$

The expression that contains ϕ and Φ, which are the density function and CDF of the standard normal distribution, respectively, is due to the truncation of ζ to the interval (0,1). Table 1 gives $\zeta^{(t)}$ from each algorithm for iterations 0–4 and 9, at which point all three algorithms had converged. For the HEM algorithm, the precision parameter $\lambda^{(t)}$ is also given. The HEM algorithm, for which the estimate of λ is approximately 16.3, estimates ζ to be about 0.6214. This lies between the original EM algorithm (0.6268) and the PEM algorithm (0.6204), the former effectively using $\lambda = 0$ and the latter fixing $\lambda = 20$. The final estimate of ζ from the HEM algorithm is relatively insensitive to the hyperpenalty. For example, if instead we choose a and b to satisfy $\mathrm{E}(\lambda) = 100$ and $\mathrm{Var}[\lambda] = 0.25$, representing a high degree of certainty about a value of λ very far from 20, the final estimates of ζ and λ are respectively 0.6019 and 99.997.

Table 1 Parameter estimates for the multinomial example from the EM algorithm and two extensions. In addition to ζ, the HEM algorithm estimates λ, a prior precision on ζ; it is set to $\lambda = 0$ and $\lambda = 20$ for the EM and PEM algorithms, respectively. In all cases, the sequence $\{\zeta^{(t)}\}$ had converged to 7 decimal places at $t = 9$.

Iteration (t)	EM ($\zeta^{(t)}$)	PEM ($\zeta^{(t)}$)	HEM ($\zeta^{(t)}$)	HEM ($\lambda^{(t)}$)
0	0.2500	0.2500	0.2500	20.0000
1	0.5576	0.5544	0.5544	16.8384
2	0.6171	0.6113	0.6122	16.4446
3	0.6255	0.6192	0.6201	16.3626
4	0.6266	0.6202	0.6212	16.3534
⋮	⋮	⋮	⋮	⋮
9	0.6268	0.6204	0.6214	16.3517

4 Example 2: Gene Expression Analysis

The multinomial example served as a proof of concept, comparing the HEM algorithm to two existing approaches; all three yielded virtually identical results. The proceeding analysis demonstrates its utility in a situation for which determining an appropriate data-adaptive value of λ is critical, and the machinery of the (P)EM algorithms is inadequate.

Chen et al. (2011) analyzed a gene-expression microarray dataset of 439 lung adenocarcinomas, originally from Shedden et al. (2008), with the goal of using gene expression plus clinical covariates – age, tumor stage and sex – to identify high-risk patients. The outcome of interest was survival after tumor resection. Expression was measured using Affymetrix oligonucleotide microarray technology. 91 promising genes were identified and re-assayed using quantitative real-time polymerase chain reaction (qRT-PCR). qRT-PCR is more precise than Affymetrix and is readily applicable in a clinical setting, thus the aim was a qRT-PCR-based prediction model. However, because of tissue availability, only 47 out of 439 tumors were re-assayed with qRT-PCR, creating a high-dimensional, missing data problem. That is, Affymetrix data was available for all 439 tumors, but the 91 qRT-PCR measurements from 392 tumors were unobserved. Chen et al. used random survival forests, together with multiple imputation of the 91×392 missing qRT-PCR measurements, to calculate a 'mortality risk index' (MRI) based on clinical covariates and qRT-PCR measurements; gene-expression measurements using Affymetrix are

not in the MRI model. Plugging MRI as a continuous covariate into a Cox model fitted to a separate validation dataset of size 100 for which qRT-PCR information was collected, the authors found good separation of the survival curves between tertiles of the MRI. More recently, Boonstra et al. (2014a) analyzed these data with a Bayesian ridge regression, using a Gibbs sampler to estimate parameters and Empirical Bayes methods to estimate the ridge parameter λ.

In our study, we use the same data to construct a model for directly predicting survival time after tumor resection by maximizing a hyperpenalized log-likelihood. Let Y denote survival time, in the logarithm of the number of months after surgery. Because some subjects are censored, say at time S, we only observe $T = \min\{Y, S\}$ and $C = 1[S < Y]$. The prediction model for Y will take as inputs the $k = 91$ qRT-PCR measurements, $\boldsymbol{X}$, and clinical covariates age, tumor stage, and sex, collectively a vector $\boldsymbol{Z}$ of length $m = 3$. We seek to use the statistical information in the 392 tumors assayed using only Affymetrix, $\boldsymbol{W}$, also of length 91 by exploiting the matched relationship of $\boldsymbol{X}$ and $\boldsymbol{W}$, namely that they are vectors of identical length, corresponding to different assays of the same 91 genes. An important question is whether the inclusion of the gene expression measurements improves predictions beyond using the clinical covariates alone.

We call the $n_{\mathrm{A}} = 47$ patients with complete covariate information *subsample A*, and the remaining $n_{\mathrm{B}} = 392$ patients are *subsample B*. The observed data from each are, respectively, $\{\boldsymbol{t}_{\mathrm{A}}, \boldsymbol{c}_{\mathrm{A}}, \boldsymbol{x}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{A}}, \boldsymbol{z}_{\mathrm{A}}\}$ and $\{\boldsymbol{t}_{\mathrm{B}}, \boldsymbol{c}_{\mathrm{B}}, \boldsymbol{w}_{\mathrm{B}}, \boldsymbol{z}_{\mathrm{B}}\}$. There

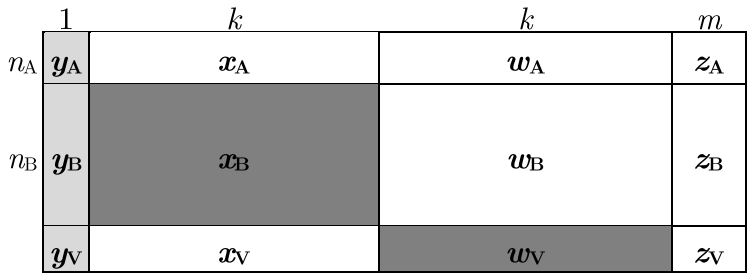

Fig. 1 Graphical representation of the training (top rows) and validation (bottom row) data. The dark-grey cells are completely missing, the light-grey cells are coarsened, meaning only a censoring time T and indicator C are observed, and the white cells are completely observed.

are two sources of missing information. First, 23 and 181 survival times are censored in subsamples A and B, respectively, and thus we observe only coarsened versions of the the vectors $\boldsymbol{y}_{\mathrm{A}}$ and $\boldsymbol{y}_{\mathrm{B}}$. Second, what is denoted as $\boldsymbol{x}_{\mathrm{B}}$, the $n_{\mathrm{B}} \times k$ matrix of qRT-PCR covariates from subsample B, is completely missing, and we only have a matched surrogate $\boldsymbol{w}_{\mathrm{B}}$ of identical dimensions. In summary, the observed data are $\boldsymbol{U}^{\mathrm{obs}} = \{\boldsymbol{t}_{\mathrm{A}}, \boldsymbol{c}_{\mathrm{A}}, \boldsymbol{x}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{A}}, \boldsymbol{z}_{\mathrm{A}}, \boldsymbol{t}_{\mathrm{B}}, \boldsymbol{c}_{\mathrm{B}}, \boldsymbol{w}_{\mathrm{B}}, \boldsymbol{z}_{\mathrm{B}}\}$ and the missing data are $\boldsymbol{U}^{\mathrm{mis}} = \{\boldsymbol{y}_{\mathrm{A}}, \boldsymbol{y}_{\mathrm{B}}, \boldsymbol{x}_{\mathrm{B}}\}$. The data are presented schematically in Figure 1.

The likelihood is constructed based upon three models:

$$
\begin{aligned}
Y|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Z} &\sim N\{\beta_0 + \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{Z}^\top \boldsymbol{\gamma}, \sigma^2\}, \\
\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Z} &\sim N_k\{\boldsymbol{\psi} + \boldsymbol{\Omega}\boldsymbol{X}, \tau^2 \boldsymbol{I}_p\}, \\
\boldsymbol{X}|\boldsymbol{Z} &\sim N_k\{\boldsymbol{\mu} + \mathbf{1}_k \boldsymbol{\delta}^\top \boldsymbol{Z}, \boldsymbol{\Sigma}\},
\end{aligned}
\tag{7}
$$

The parameters β_0, σ^2, and τ^2 are scalar-valued. $\boldsymbol{\beta}$, $\boldsymbol{\psi}$ and $\boldsymbol{\mu}$ are length-k. $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are length-m, corresponding to effect of age, tumor stage, and sex on Y

and the components of $\boldsymbol{X}$, respectively. $\boldsymbol{\Omega}$ is a diagonal $k \times k$ matrix and $\boldsymbol{\Sigma}$ is an unconstrained $k \times k$ covariance matrix. That $\boldsymbol{\Omega}$ is diagonal implicitly assumes that the Affymetrix measurement of the ith gene depends only on the qRT-PCR measurement of the ith gene and not that of the jth gene, $j \neq i$. The final model above implies that, marginally for gene j, $X_j|\boldsymbol{Z} \sim N\{\mu_j + \boldsymbol{\delta}^\top \boldsymbol{Z}, \Sigma_{jj}\}$.

We also highlight two conditional independence assumptions: (i) Y and $\boldsymbol{W}$ are conditionally independent given $\boldsymbol{X}$ and $\boldsymbol{Z}$ and (ii) $\boldsymbol{W}$ and $\boldsymbol{Z}$ are conditionally independent given $\boldsymbol{X}$. Finally, we assume that the distribution of the censoring indicator C does not depend on any components of $\boldsymbol{\theta}$.

We consider four EM-type algorithms with increasing amounts of penalization. The first, called option 1, uses both clinical ($\boldsymbol{Z}$) and genomic ($\boldsymbol{X}$) covariates but does not shrink any corresponding regression coefficients (as discussed later, mild penalization of the nuisance parameter $\boldsymbol{\Sigma}^{-1}$ is required to make the algorithm proceed). The censored survival times and missing expression measurements are missing data in the E-step. At the other extreme, option 4 uses only clinical covariates ($\boldsymbol{Z}$). Although this is technically an EM algorithm with no penalization, it represents the limiting case of fitting model (7) with an increasing amount of penalization on $\boldsymbol{\beta}$, i.e. $\boldsymbol{\beta}$ set exactly to $\mathbf{0}_k$. The only missing data here are the censored outcomes Y, because $\boldsymbol{X}$ is not used. Between these extremes, we consider an HEM algorithm that includes $\boldsymbol{Z}$ and $\boldsymbol{X}$ but adaptively shrinks $\boldsymbol{\beta}$, which is the genomic contribution to the model, by a hyperpenalty. Two choices of hyperpenalty based on the gamma and inverse-gamma distributions are considered. Either of these HEM algorithms

can make superior predictions in the validation data. These four options are summarized in Table 2 and discussed as follows.

4.1 E-step

Let $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\psi}, \boldsymbol{\Omega}, \tau^2, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\Sigma}\}$ denote all of the model parameters. Up to an additive constant, the *complete-data* log-likelihood is given by

$$
\begin{aligned}
\ell_C &= \ln[\boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}} | \boldsymbol{\theta}] \\
&= \ln[\boldsymbol{y}_{\text{A}} | \boldsymbol{x}_{\text{A}}, \boldsymbol{z}_{\text{A}}, \beta_0, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2] + \ln[\boldsymbol{w}_{\text{A}} | \boldsymbol{x}_{\text{A}}, \boldsymbol{\psi}, \boldsymbol{\Omega}, \tau^2] + \ln[\boldsymbol{x}_{\text{A}} | \boldsymbol{z}_{\text{A}}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\Sigma}] \\
&\quad + \ln[\boldsymbol{y}_{\text{B}} | \boldsymbol{x}_{\text{B}}, \boldsymbol{z}_{\text{B}}, \beta_0, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2] + \ln[\boldsymbol{w}_{\text{B}} | \boldsymbol{x}_{\text{B}}, \boldsymbol{\psi}, \boldsymbol{\Omega}, \tau^2] + \ln[\boldsymbol{x}_{\text{B}} | \boldsymbol{z}_{\text{B}}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\Sigma}] \\
&= -\frac{n_{\text{A}} + n_{\text{B}}}{2} \ln(\sigma^2 \tau^{2p} / |\boldsymbol{\Sigma}^{-1}|) \\
&\quad - \frac{1}{2\sigma^2} (\boldsymbol{y}_{\text{A}} - \beta_0 \mathbf{1}_{n_{\text{A}}} - \boldsymbol{x}_{\text{A}} \boldsymbol{\beta} - \boldsymbol{z}_{\text{A}} \boldsymbol{\gamma})^\top (\boldsymbol{y}_{\text{A}} - \beta_0 \mathbf{1}_{n_{\text{A}}} - \boldsymbol{x}_{\text{A}} \boldsymbol{\beta} - \boldsymbol{z}_{\text{A}} \boldsymbol{\gamma}) \\
&\quad - \frac{1}{2} \sigma^2 (\boldsymbol{y}_{\text{B}} - \beta_0 \mathbf{1}_{n_{\text{B}}} - \boldsymbol{x}_{\text{B}} \boldsymbol{\beta} - \boldsymbol{z}_{\text{B}} \boldsymbol{\gamma})^\top (\boldsymbol{y}_{\text{B}} - \beta_0 \mathbf{1}_{n_{\text{B}}} - \boldsymbol{x}_{\text{B}} \boldsymbol{\beta} - \boldsymbol{z}_{\text{B}} \boldsymbol{\gamma}) \\
&\quad - \frac{1}{2\tau^2} \text{Tr}\ (\boldsymbol{w}_{\text{A}} - \mathbf{1}_{n_{\text{A}}} \boldsymbol{\psi}^\top - \boldsymbol{x}_{\text{A}} \boldsymbol{\Omega})^\top (\boldsymbol{w}_{\text{A}} - \mathbf{1}_{n_{\text{A}}} \boldsymbol{\psi}^\top - \boldsymbol{x}_{\text{A}} \boldsymbol{\Omega}) \\
&\quad - \frac{1}{2\tau^2} \text{Tr}\ (\boldsymbol{w}_{\text{B}} - \mathbf{1}_{n_{\text{B}}} \boldsymbol{\psi}^\top - \boldsymbol{x}_{\text{B}} \boldsymbol{\Omega})^\top (\boldsymbol{w}_{\text{B}} - \mathbf{1}_{n_{\text{B}}} \boldsymbol{\psi}^\top - \boldsymbol{x}_{\text{B}} \boldsymbol{\Omega}) \\
&\quad - \frac{1}{2} \text{Tr}\ (\boldsymbol{x}_{\text{A}} - \mathbf{1}_{n_{\text{A}}} \boldsymbol{\mu}^\top - \boldsymbol{z}_{\text{A}} \boldsymbol{\delta} \mathbf{1}_p^\top) \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_{\text{A}} - \mathbf{1}_{n_{\text{A}}} \boldsymbol{\mu}^\top - \boldsymbol{z}_{\text{A}} \boldsymbol{\delta} \mathbf{1}_p^\top)^\top \\
&\quad - \frac{1}{2} \text{Tr}\ (\boldsymbol{x}_{\text{B}} - \mathbf{1}_{n_{\text{B}}} \boldsymbol{\mu}^\top - \boldsymbol{z}_{\text{B}} \boldsymbol{\delta} \mathbf{1}_p^\top) \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_{\text{B}} - \mathbf{1}_{n_{\text{B}}} \boldsymbol{\mu}^\top - \boldsymbol{z}_{\text{B}} \boldsymbol{\delta} \mathbf{1}_p^\top)^\top. \qquad (8)
\end{aligned}
$$

The E-step would calculate $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) = \text{E}(\ell_C | \boldsymbol{U}^{\text{obs}}, \boldsymbol{\theta}^{(t)})$, where $\boldsymbol{\theta}^{(t)}$ is a current estimate of $\boldsymbol{\theta}$. However, the conditional joint distribution of $\boldsymbol{U}^{\text{mis}} = \{\boldsymbol{y}_{\text{A}}, \boldsymbol{y}_{\text{B}}, \boldsymbol{x}_{\text{B}}\}$ is intractable in closed form. We make use of a nested E-step (Van Dyk, 2000), which is described in detail in the Web Appendix C. Briefly, the missing data are partitioned into $\boldsymbol{U}^{\text{mis1}} = \{\boldsymbol{y}_{\text{B}}\}$ and $\boldsymbol{U}^{\text{mis2}} = \{\boldsymbol{y}_{\text{A}}, \boldsymbol{x}_{\text{B}}\}$.

A Gibbs sampler first calculates Monte Carlo estimates of the conditional expectations of $\boldsymbol{y}_{\mathrm{B}}$ and $\boldsymbol{y}_{\mathrm{B}}^{\top}\boldsymbol{y}_{\mathrm{B}}$, which are the sufficient statistics for $\boldsymbol{U}^{\mathrm{mis1}}$, given $\boldsymbol{U}^{\mathrm{obs}}$ and $\boldsymbol{\theta}^{(t)}$. Next, conditioning on these estimated sufficient statistics, that is, treating them as additional "observed data", run several iterations of an inner EM algorithm, now considering only $\boldsymbol{U}^{\mathrm{mis2}}$ as missing data. In other words, iterate between calculating expectations of the sufficient statistics for $\boldsymbol{U}^{\mathrm{mis2}}$, these being $\boldsymbol{y}_{\mathrm{A}}$, $\boldsymbol{y}_{\mathrm{A}}^{\top}\boldsymbol{y}_{\mathrm{A}}$, $\boldsymbol{x}_{\mathrm{B}}$ and $\boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{x}_{\mathrm{B}}$, and updating $\boldsymbol{\theta}$ (and $\boldsymbol{\eta}$, when necessary), all the while conditioning on the sufficient statistics for $\boldsymbol{U}^{\mathrm{mis1}}$. By nesting the E-steps, we avoid the need for the joint distribution $\boldsymbol{U}^{\mathrm{mis1}}$ and $\boldsymbol{U}^{\mathrm{mis2}}$. Moreover, this is computationally fast because the E-step in the inner loop is available in closed form.

Remark 4 The E-step for the algorithm that uses only the clinical data, option 4 in Table 2, is different and considerably simpler than above, because only the sufficient statistics for $\boldsymbol{y}_{\mathrm{A}}$ and $\boldsymbol{y}_{\mathrm{B}}$ need to be calculated at each iteration, and no nested E-step is required.

4.2 M-Steps

Recalling the components of the general form of the M- and H-steps in (2) and (3), we consider two choices of penalty $p_{\boldsymbol{\eta}}(\boldsymbol{\theta})$. The first penalty, which is option 1 in Table 2, fixes $\boldsymbol{\eta}$ at a given value and therefore does not require a hyperpenalty $h(\boldsymbol{\eta})$. The second choice of $p_{\boldsymbol{\eta}}(\boldsymbol{\theta})$, options 2–3 in Table 2, allows $\boldsymbol{\eta}$ to be unknown and results in data-adaptive shrinkage (option 4 does not make

Table 2 Summary of four options considered, ordered by increasing amounts of shrinkage of $\boldsymbol{\beta}$, the regression coefficients in (7). Option 4 does not use $\boldsymbol{X}$ and therefore is equivalently interpreted as setting $\boldsymbol{\beta}$ to $\mathbf{0}_k$; technically, however, it is an EM algorithm and has no penalization.

Option	Algorithm	Covariates Used	Description of Penalty (Equation)	Hyperpenalty (Equation)
1	PEM	$\boldsymbol{X}$,$\boldsymbol{Z}$	mild shrinkage of $\boldsymbol{\Sigma}^{-1}$, no shrinkage of $\boldsymbol{\beta}$, (9)	none
2	HEM	$\boldsymbol{X}$,$\boldsymbol{Z}$	mild shrinkage of $\boldsymbol{\Sigma}^{-1}$, adaptive shrinkage of $\boldsymbol{\beta}$, (10)	gamma (11)
3	HEM	$\boldsymbol{X}$,$\boldsymbol{Z}$	mild shrinkage of $\boldsymbol{\Sigma}^{-1}$, adaptive shrinkage of $\boldsymbol{\beta}$, (10)	inverse-gamma (12)
4	EM	$\boldsymbol{Z}$	$\boldsymbol{\beta}$ set exactly equal to $\mathbf{0}_k$	none

use of a penalty). In this latter case, a hyperpenalty $h(\boldsymbol{\eta})$ must be selected, and we consider two choices that both satisfy the result in Section 2.

Option 1: PEM Maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ may result in an infinite likelihood when neither subsamples A nor B contains enough information about $\boldsymbol{\Sigma}^{-1}$. From preliminary simulation studies based on this data structure, this will happen when $k > n_{\mathrm{A}}$, so that the sample covariance of $\boldsymbol{x}_{\mathrm{A}}$ is singular and, simultaneously, when τ^2 is large relative to the elements of $\boldsymbol{\Omega}$, meaning the surrogate $\boldsymbol{w}_{\mathrm{B}}$ is too noisy to provide information about $\boldsymbol{\Sigma}^{-1}$. In this case, the estimate for $\boldsymbol{\Sigma}^{-1}$ approaches singularity and the expected log-likelihood diverges. To address this problem, we mildly penalize the estimate of $\boldsymbol{\Sigma}^{-1}$:

$$p_{\boldsymbol{\eta}}(\boldsymbol{\theta}) \equiv p(\boldsymbol{\theta}) = \frac{2k-1}{2} \ln |\boldsymbol{\Sigma}^{-1}| - \frac{2k-1}{2} \mathrm{Tr}\ (\mathrm{diag}(\hat{\mathrm{Var}}[\boldsymbol{x}_{\mathrm{A}}])\boldsymbol{\Sigma}^{-1}). \quad (9)$$

The matrix $\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])$ is a $k \times k$ matrix containing the diagonal elements of the sample variance of $\boldsymbol{x}_\text{A}$. This penalty corresponds to a Wishart prior on $\boldsymbol{\Sigma}^{-1}$ with $3k$ degrees of freedom and scale matrix $\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])$. The penalty parameter $\boldsymbol{\eta}$ is pre-selected, so this is a PEM algorithm. This small amount of shrinkage induced by $p(\boldsymbol{\theta})$ is sufficient to make the algorithm proceed without an infinite log-likelihood and is related to but different than shrinkage of $\boldsymbol{\beta}$. Following Meng and Rubin (1993), we divide $\boldsymbol{\theta}$ into sub-components and use conditional penalized M-steps to update each sub-component individually; these are derived in Web Appendices D & E .

Options 2 & 3: HEM In both Options 2 and 3, the penalty function is expanded relative to (9) as follows:

$$\begin{aligned} p_{\boldsymbol{\eta}}(\boldsymbol{\theta}) = &\frac{2k-1}{2} \ln |\boldsymbol{\Sigma}^{-1}| - \frac{2k-1}{2} \text{Tr } (\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])\boldsymbol{\Sigma}^{-1}) \\ &- \frac{k}{2} \ln(\sigma^2) + \frac{k}{2} \ln(\lambda) - \frac{1}{2\sigma^2} \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}. \end{aligned} \tag{10}$$

This adds a normal log-density term to the penalty function from option 1 in (9) and so corresponds to a ridge regression. Now, $p_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ contains a tuning parameter λ, that is, $\boldsymbol{\eta} = \{\lambda\}$, for which a value must be chosen. We adaptively choose λ using a H-step as in (2), with two choices of the hyperpenalty $h(\lambda)$ based on the gamma and inverse-gamma distributions. Each is indexed by shape and rate parameters, respectively a and b:

$$h_{\text{gamma}}(\lambda) = -(a-1)\ln(\lambda) + \lambda/b, \tag{11}$$

$$h_{\text{inv-gamma}}(\lambda) = (a+1)\ln(\lambda) + 1/(b\lambda). \tag{12}$$

Based on discussion in Boonstra et al. (2014b), for each case we chose a and b to satisfy the moment-matching conditions $\mathrm{E}(\lambda) = k(1/\tilde{R}^2 - 1) = 364$ and $\mathrm{Var}[\lambda] = 2k(1/\tilde{R}^2 - 1)^2 = 2912$, where k is the length of $\boldsymbol{\beta}$ and $\tilde{R}^2 = 0.2$ is our prior guess at the coefficient of determination, or proportion of variance explained by the covariates, in (7). For the gamma hyperpenalty, this gives $a = k/2$ and $b = (1/\tilde{R}^2 - 1)^{-1}/2$. For the inverse-gamma hyperpenalty, this gives $a = k/2 + 2$ and $b = (1/\tilde{R}^2 - 1)^{-1}/(k[k/2 + 1])$.

Remark 5 The strategy of setting $\mathrm{E}(\lambda) = k(1/\tilde{R}^2 - 1)$ above is based on the result in Hoerl et al. (1975), who show that, when fitting a linear model with uncorrelated covariates, the optimal λ in terms of mean squared prediction error of $\boldsymbol{\beta}$ is $\lambda^* = k\sigma^2/\boldsymbol{\beta}^\top\boldsymbol{\beta} = k(1/R^2 - 1)$. There is no such closed form when the covariates arise according to an arbitrary correlation structure, but results in Boonstra et al. (2014b) suggest that this strategy can still choose λ close to the optimal value under various correlation structures, even if $\tilde{R}^2 \neq R^2$.

Remark 6 The gamma hyperpenalty satisfies the finite HLL result from section 2 for any $a, b > 0$. For the inverse-gamma hyperpenalty, the condition $a > k/2 - 1$ ensures a finite HLL, which is satisfied here.

The hyperpenalized M-steps are respectively as follows:

$$\text{HEM(GA)}: \quad \lambda^{(t+1)} = \frac{k + 2a - 2}{\boldsymbol{\beta}^{(t)\top}\boldsymbol{\beta}^{(t)}/\sigma^{2(t)} + 2b},$$

$$\text{HEM(IG)}: \quad \lambda^{(t+1)} = \frac{k - 2a - 2 + \sqrt{(k - 2a - 2)^2 + 8\boldsymbol{\beta}^{(t)\top}\boldsymbol{\beta}^{(t)}/(b\sigma^{2(t)})}}{2\boldsymbol{\beta}^{(t)\top}\boldsymbol{\beta}^{(t)}/\sigma^{2(t)}}.$$

In summary, we consider a PEM algorithm that uses the genomic data but does not shrink its contribution, two implementations of the HEM algorithm that adaptively shrink the contribution from the genomic data, and an EM algorithm that fits only the clinical covariates.

The results from fitting each model are in Figure 2 and Table 3. The HEM(GA) and HEM(IG) methods selected $\lambda = 641$ and $\lambda = 1539$, respectively. In the figure, individuals in the validation data are separated based on whether each predicted survival time is less than 30 months, between 30 and 60 months, or longer than 60 months, and we compare the three Kaplan-Meier (KM) curves. Visually, option 1 best separates the curves, followed by options 2 and 3, in which the two higher-risk groups are less distinguishable early on. For option 4, the curves cross several times. The estimated median survival times for the "less than 30 month" risk group was 25.5, 28.6, 28.6, and 32.3 months for options 1–4, respectively. For the "30 to 60 month" risk group, these median survival times were 67.9, 67.9, 46.3, and 31.1 months. This indicates that option 4 can not distinguish between the high- and medium-risk groups. The median was not crossed for the "longer than 60 month" group. Visually, this group had worse survival in options 2 and 4.

Included in Table 3 is the range of each $\boldsymbol{\beta}$. The HEM algorithms typically shrink $\boldsymbol{\beta}$ by a factor of 3 to 5, relative to the PEM algorithm. Table 3 contains an estimate of the concordance probability (c, Harrell, 2001), which is the proportion of all possible pairs of individuals for which the ordering of predicted survival times matches that of the actual survival times; a larger c is better.

We calculated c in both the training (c_{train}) and validation (c_{validate}) data. Option 1 has the best c_{train} (0.889), but options 2–4 give better values of c_{validate} (respectively 0.696, 0.698, and 0.681) compared to option 1 (0.671). Also given is each option's scaled integrated brier score (SIBS, Graf et al., 1999; Peters and Hothorn, 2013) for the validation data, which is an overall measure of predictive accuracy that accounts for censoring. The scores are scaled so that a SIBS of 1 indicates a purely equivocal model, i.e. a predicted survivor function of 0.5 at all times for any person; a smaller SIBS is better. Options 2 and 3 have the best SIBS (both 0.359), followed by option 4 (0.374) and then option 1 (0.382). To summarize, the HEM algorithms are overall well-calibrated, provide good discrimination between estimated risk groups, and have the smallest prediction error, relative to either including the genomic information but not shrinking its contribution (option 1) or not including the genomic information at all (option 4).

5 Discussion

The HEM algorithm is a natural extension to the PEM algorithm for scenarios in which the penalty parameter $\boldsymbol{\eta}$ is not known. Rather, support for a plausible range of values is provided through a hyperpenalty, $h(\boldsymbol{\eta})$, and the HEM algorithm simultaneously chooses both the main inferential parameter, $\boldsymbol{\theta}$, and the penalty parameter, $\boldsymbol{\eta}$, that best match the data and the hyperpenalty. The hyperpenalty corresponds to a hyperprior in a Bayesian analysis, but, as

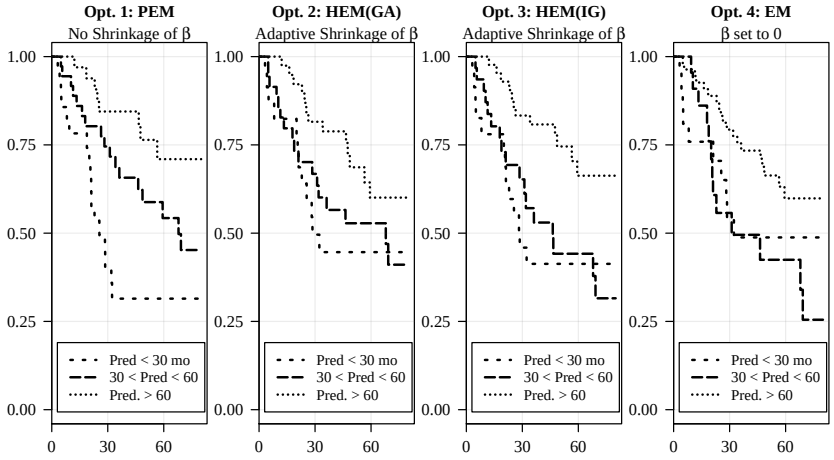

Fig. 2 Results from the analysis of gene expression data applied to validation data. The PEM (left) includes the genomic information but does not shrink its contribution. The EM (right) does not include the genomic information, effectively setting its contribution to zero and therefore corresponding to a large amount of penalization. Between, λ is adaptively estimated with two different hyperpenalty functions.

demonstrated in the simple case of ridge regression, not all hyperpenalties, even if they are proper, guarantee that the maximum HLL will be finite. We established a simple sufficient condition on the hyperpenalty to ensure a finite maximum HLL.

The HEM algorithm can be used in hierarchical analyses with no actual missing data but for which some parameters are treated as "missing data" in the E-step, as in Yi and Xu (2008) or Mutshinda and Sillanpää (2012). In these scenarios, a hyperpenalty would adaptively shrink parameters at the bottom of the hierarchy, for example the degrees of freedom parameter of a student-t prior. Here, we applied the HEM algorithm to two examples with actual

Table 3 Numerical results from the analysis of gene expression data. The first rows give the value of the shrinkage parameter λ and the range of the resulting $\boldsymbol{\beta}$'s. The c's are estimates of the concordance probabilities applied to both the training and validation data (larger is better), and SIBS is the scaled integrated brier score (smaller is better). In contrast to Figure 2, option 1, which does not shrink the contribution from the genomic information, is least-preferred.

	Opt. 1	Opt. 2	Opt. 3	Opt. 4
	PEM	HEM(GA)	HEM(IG)	EM
λ	0	641	1539	∞
range($\boldsymbol{\beta}$)	(-0.049, 0.054)	(-0.017, 0.012)	(-0.009, 0.007)	(0,0)
c_{train}	0.889	0.849	0.826	0.796
c_{validate}	0.671	0.696	0.698	0.681
SIBS	0.382	0.359	0.359	0.374

missing data, the first being a simple demonstrative application. The second was a more complex analysis trying to incorporate gene expression data to better predict survival time in lung cancer patients. Owing to the dimension of the problem, shrinkage of the regression coefficients was desired, but cross-validation-type methods, the typical approach for selecting such parameters, are not easily applied in a missing data context such as this. Adaptive shrinkage of $\boldsymbol{\beta}$ via the HEM algorithm provided good separation of risk curves and better overall measures of discrimination and predictive ability relative to both no shrinkage at all and not using any of the genomic information.

Acknowledgements This work was supported by the National Science Foundation [DMS1007494] and the National Institutes of Health [CA129102]. Code for both data analysis examples was written in `R` (R Core Team, 2014) and is available at `http://www-personal.umich.edu/~philb`.

Conflict of interest: none (PSB, JMGT, and BM).

References

Boonstra, P. S., Mukherjee, B., and Taylor, J. M. G. (2014a). Bayesian shrinkage methods for partially observed data with many predictors. *Annals of Applied Statistics* **14,** 2272–2292.

Boonstra, P. S., Mukherjee, B., and Taylor, J. M. G. (2014b). A small-sample choice of the tuning parameter in ridge regression. *Statistica Sinica,* In Press.

Casella, G. (2001). Empirical Bayes Gibbs sampling. *Biostatistics* **2,** 485–500.

Chen, G., Kim, S., Taylor, J. M. G., Wang, Z., Lee, O., Ramnath, N., Reddy, R. M., Lin, J., Chang, A. C., Orringer, M. B., and Beer, D. G. (2011). Development and validation of a qRT-PCR–classifier for lung cancer prognosis. *Journal of Thoracic Oncology* **6,** 1481–1487.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* **39,** 1–38.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85,** 398–409.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18,** 2529–2545.

Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society: Series B* **52,** 443–452.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis.* Springer, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning.* Springer, New York, 2nd edition.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12,** 55–67.

Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics-Theory and Methods* **4,** 105–123.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80,** 267–278.

Mutshinda, C. M. and Sillanpää, M. J. (2012). Swift block-updating EM and pseudo-EM procedures for Bayesian shrinkage analysis of quantitative trait loci. *Theoretical and Applied Genetics* **125,** 1575–1587.

Peters, A. and Hothorn, T. (2013). *ipred: Improved Predictors.* R package version 0.9-2.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Shedden, K. et al. (2008). Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study. *Nature Medicine* **14,** 822–827.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82,** 528–540.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B* **58,** 267–288.

Van Dyk, D. (2000). Nesting EM algorithms for computational efficiency. *Statistica Sinica* **10,** 203–226.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the Poor Man's Data Augmentation algorithms. *Journal of the American Statistical Association* **85,** 699–704.

Wu, C. F. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11,** 95–103.

Yi, N. and Xu, S. (2008). Bayesian Lasso for quantitative trait loci mapping. *Genetics* **179,** 1045–1055.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67,** 301–320.