Analytic solution and stationary phase approximation for the Bayesian lasso and elastic net

ANALYTIC SOLUTION AND STATIONARY PHASE APPROXIMATION FOR THE BAYESIAN LASSO AND ELASTIC NET

TOM MICHOEL

ABSTRACT. Regression shrinkage and variable selection are important concepts in high-dimensional statistics that allow the inference of robust models from large data sets. Bayesian methods achieve this by subjecting the model parameters to a prior distribution whose mass is centred around zero. In particular, the lasso and elastic net linear regression models employ a double-exponential distribution in their prior, which results in some maximum-likelihood regression coefficients being identically zero. Because of their ability to simultaneously perform parameter estimation and variable selection, these models have become enormously popular. However, there has been limited success in moving beyond maximum-likelihood estimation and deriving estimates for the posterior distribution of regression coefficients, due to a need for computationally expensive Gibbs sampling approaches to evaluate analytically intractable partition function integrals. Here, through the use of the Fourier transform, these integrals are expressed as complex-valued oscillatory integrals over "regression frequencies". This results in an analytic expansion and stationary phase approximation for the partition functions of the Bayesian lasso and elastic net, where the non-differentiability of the double-exponential prior distribution has so far eluded such an approach. Use of this approximation leads to highly accurate numerical estimates for the expectation values and marginal posterior distributions of the regression coefficients, thus allowing for Bayesian inference of much higher dimensional models than previously possible.

1. INTRODUCTION

Modern statistical modelling and inference involves high-dimensional data sets where the number of variables far exceeds the number of experimental samples. Application of traditional regression methods typically results in over-fitted models that do not generalize well to unseen data. Prediction accuracy in these situations can often be improved by shrinking regression coefficients towards zero, or setting some of them equal to zero (Friedman et al., 2001). Bayesian methods achieve this by performing an ordinary regression subject to a prior distribution on the regression coefficients whose mass is concentrated around zero. For least squares regression, the most popular methods are ridge regression (Hoerl and Kennard, 1970), corresponding to a normally distributed prior; lasso regression (Tibshirani, 1996), corresponding to a double-exponential or Laplace distribution prior; and elastic net regression (Zou and Hastie, 2005), whose prior interpolates between the lasso and ridge priors. Of these, only the lasso and elastic net result in a selection of variables, i.e. in their maximum-likelihood solutions, a subset of regression coefficients are exactly zero.

Although the maximum-likelihood lasso and elastic net regression models have proven extremely powerful across a wide range of application domains, they only provide a point estimate for the regression coefficients. A fully Bayesian treatment that takes into account uncertainty due to data noise and limited sample size, and provides posterior distributions and confidence intervals, is therefore of great interest. Unsurprisingly, Bayesian inference for the lasso and elastic net involves analytically intractable integrals and requires the use of numerical Gibbs sampling techniques (Park and Casella, 2008; Hans, 2009; Li et al., 2010; Hans, 2011). However, Gibbs sampling is computationally expensive and, particularly in high-dimensional settings, convergence may be slow and difficult to assess or remedy (Liu, 2004; Mallick and Yi, 2013; Rajaratnam and Sparks, 2015a,b). An alternative to Gibbs sampling for Bayesian inference is to use asymptotic approximations to the intractable integrals based on Laplace's method (Kass and Steffey, 1989;

Research supported by the BBSRC (grant numbers BB/J004235/1 and BB/M020053/1).

Rue et al., 2009). However, the log-likelihoods of the lasso and elastic net models contain a non-differentiable term proportional to the ℓ_1-norm (i.e. sum of absolute values) of the regression coefficients, and are therefore off-limits to the Laplace approximation which requires twice differentiable log-likelihood functions.

The aim of this paper is to show that approximate Bayesian inference is in fact possible using a Laplace-like approximation, more precisely the stationary phase or saddle point approximation for complex-valued oscillatory integrals (Wong, 2001). This is achieved by rewriting the integrals in question as a function of "frequencies" instead of as a function of the regression coefficients, through the use of the Fourier transform. The appearance of the Fourier transform in this context should not come as an altogether big surprise. The stationary phase approximation can be used to obtain or invert characteristic functions, which are of course Fourier transforms (Daniels, 1954). More to the point of this paper, there is an intimate connection between the Fourier transform of the exponential of a convex function and the Legendre-Fenchel transform of that convex function, which plays a fundamental role in physics by linking microscopic statistical mechanics to macroscopic thermodynamics, or quantum to classical mechanics (Litvinov, 2005). In particular, convex duality (Boyd and Vandenberghe, 2004; Rockafellar, 1970), which maps the solution of a convex optimization problem to that of its dual, is essentially equivalent to writing the partition function of a Gibbs probability distribution in coordinate or frequency space (Appendix A).

Convex duality principles have been essential to characterize analytical properties of the maximum-likelihood solutions of the lasso and elastic net regression models (Osborne et al., 2000a,b; El Ghaoui et al., 2012; Tibshirani et al., 2012; Tibshirani, 2013; Michoel, 2016). This paper shows that equally powerful duality principles exist to study Bayesian inference problems.

2. Analytic results

We consider the usual setup for linear regression where there are n observations of p predictor variables and one response variable, and the effects of the predictors on the response are to be determined by minimizing the least squares cost function $\|y - Ax\|^2$ subject to additional constraints, where $y \in \mathbb{R}^n$ are the response data, $A \in \mathbb{R}^{n\times p}$ are the predictor data, $x \in \mathbb{R}^p$ are the regression coefficients which need to be estimated and $\|v\| = (\sum_{i=1}^n |v_i|^2)^{1/2}$ is the ℓ^2-norm. Without loss of generality, it is assumed that the response and predictors are centred and standardized,

$$\sum_{i=1}^n y_i = \sum_{i=1}^n A_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^n y_i^2 = \sum_{i=1}^n A_{ij}^2 = n \quad \text{for } j \in \{1,2,\dots,p\}. \tag{1}$$

In a Bayesian setting, a hierarchical model is assumed where each sample y_i is drawn independently from a normal distribution with mean $A_{i\bullet}x$ and variance σ^2, where $A_{i\bullet}$ denotes the i^{th} row of A, or more succintly,

$$y \mid A, x \sim \mathcal{N}(Ax, \sigma^2 \mathbb{1}), \tag{2}$$

where $\mathcal{N}$ denotes a multivariate normal distribution, and the regression coefficients x are assumed to have a prior distribution

$$x \sim \exp\Big[-\frac{n}{\sigma^2}\big(\lambda\|x\|^2 + 2\mu\|x\|_1\big)\Big], \tag{3}$$

where $\|x\|_1 = \sum_{j=1}^p |x_j|$ is the ℓ^1-norm, and the prior distribution is defined upto a normalization constant. The apparent dependence of the prior distribution on the data via the dimension paramater n only serves to simplify notation, allowing the posterior distribution of the regression coefficients to be written, using Bayes' theorem, as

$$p(x \mid y, A) \propto p(y \mid x, A)p(x) \propto e^{-\frac{n}{\sigma^2}\mathcal{L}(x)}, \tag{4}$$

where

$$\mathcal{L}(x) = \frac{1}{2n}\|y - Ax\|^2 + \lambda\|x\|^2 + 2\mu\|x\|_1 \tag{5}$$

$$= x^T\big(\frac{A^TA}{2n} + \lambda\mathbb{1}\big)x - 2\big(\frac{A^Ty}{2n}\big)^T x + 2\mu\|x\|_1 + \frac{1}{2n}\|y\|^2 \tag{6}$$

is minus the posterior log-likelihood function. The maximum-likelihood solutions of the lasso ($\lambda = 0$) and elastic net ($\lambda > 0$) models are obtained by minimizing $\mathcal{L}$, where the relative scaling of the penalty parameters to the sample size n corresponds to the notational conventions of Friedman et al. (2010)*. In the current setup, it is assumed that the parameters $\lambda \geq 0$, $\mu > 0$ and $\sigma^2 > 0$ are given a priori.

To facilitate notation, a slightly more general class of cost functions is defined as

$$H(x \mid C, w, \mu) = x^TCx - 2w^Tx + 2\mu\|x\|_1, \tag{7}$$

where $C \in \mathbb{R}^{p\times p}$ is a positive-definite matrix, $w \in \mathbb{R}^p$ is an arbitrary vector and $\mu > 0$. After discarding the constant term $\frac{1}{2n}\|y\|^2$, $\mathcal{L}(x)$ is of this form, as is the so-called "non-naive" elastic net, where $C = \frac{\frac{1}{2n}A^TA+\lambda\mathbb{1}}{\lambda+1}$ (Zou and Hastie, 2005). More importantly perhaps, eq. (7) also covers linear mixed models, where samples need not be independent (Rakitsch et al., 2012). In this case, eq. (2) is replaced by

$$y \mid A, x \sim \mathcal{N}(Ax, \sigma^2K),$$

for some covariance matrix $K \in \mathbb{R}^{n\times n}$, resulting in a posterior minus log-likelihood function with $C = \frac{1}{2n}A^TK^{-1}A + \lambda\mathbb{1}$ and $w = \frac{1}{2n}A^TK^{-1}y$.

The requirement that C is positive definite, and hence invertible, implies that H is strictly convex and hence has a unique minimizer. For the lasso ($\lambda = 0$) this only holds without further assumptions if $n \geq p$ (Tibshirani, 2013); for the elastic net ($\lambda > 0$) there is no such constraint. The Gibbs distribution on $\mathbb{R}^p$ for the cost function $H(x \mid C, w, \mu)$ with inverse temperature τ is defined as

$$p(x) = \frac{e^{-\tau H(x)}}{Z},$$

where for ease of notation we have dropped explicit reference to C, w and μ. The normalization constant $Z = \int_{\mathbb{R}^p} e^{-\tau H(x)}dx$ is called the partition function. There is no known analytic solution for the partition function integral. However, in the posterior distribution (4), the inverse temperature $\tau = \frac{n}{\sigma^2}$ is large, firstly because we are interested in high-dimensional problems where n is large (even if it may be small compared to p), and secondly because we assume a priori that (some of) the predictors are informative for the response variable and that therefore σ^2, the amount of variance of y unexplained by the predictors, must be small.

It therefore makes sense to seek an analytic approximation to the partition function for large values of τ. However, the usual approach to approximate $e^{-\tau H(x)}$ by a Gaussian in the vicinity of the minimizer of H and apply a Laplace approximation (Wong, 2001) is not feasible, because H is not twice differentiable. Instead we observe that $e^{-\tau H(x)} = e^{-2\tau f(x)}e^{-2\tau g(x)}$ where

$$f(x) = \frac{1}{2}x^TCx - w^Tx \tag{8}$$

$$g(x) = \mu\sum_{j=1}^{p}|x_j|. \tag{9}$$

Using Parseval's identity for Fourier transforms (Appendix A.1), it follows that (Appendix A.3)

$$Z = \int_{\mathbb{R}^p} e^{-2\tau f(x)}e^{-2\tau g(x)}dx = \frac{\mu^p}{(\pi\tau)^{\frac{p}{2}}\sqrt{\det(C)}}\int_{\mathbb{R}^p}\frac{e^{-\tau(k-iw)^TC^{-1}(k-iw)}}{\prod_{j=1}^{p}(k_j^2+\mu^2)}dk. \tag{10}$$

*To be precise, Friedman et al. (2010) write the penalty term as $\tilde{\lambda}(\frac{1-\alpha}{2}\|x\|_2^2 + \alpha\|x\|_1)$, wich is obtained from (5) by setting $\tilde{\lambda} = 2(\lambda+\mu)$ and $\alpha = \frac{\mu}{\lambda+\mu}$.

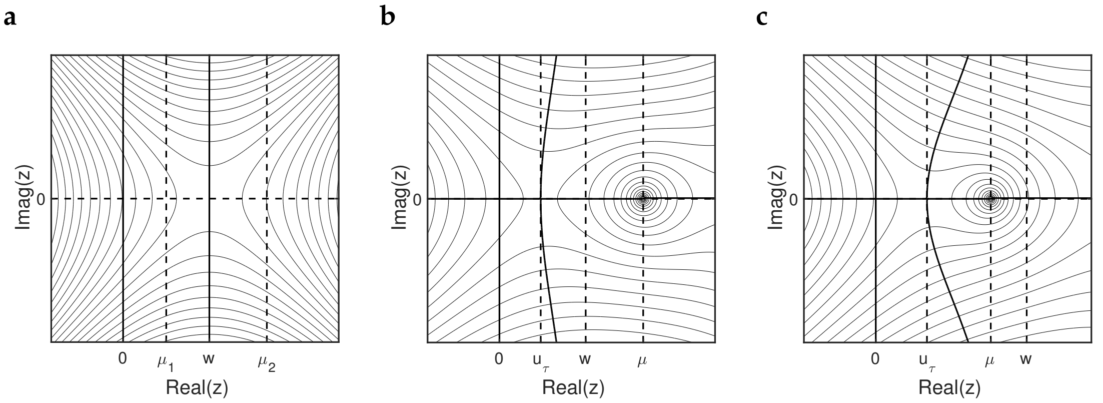

FIGURE 1. Illustration of the stationary phase approximation procedure for $p = 1$. **(a)** Contour plot of the complex function $(z - w)^2$. If $\mu = \mu_2$, the integration contour can be deformed from the imaginary axis to a steepest descent contour parallel to the imaginary axis and passing through the saddle point $z_0 = w$, whereas if $\mu = \mu_1$, this cannot be done without passing through the pole at $z = \mu$. **(b,c)** Contour plots of the complex function $(z - w)^2 - \frac{1}{\tau}\ln(\mu^2 - z^2)$ for $|w| < \mu$ and $|w| \geq \mu$, respectively. In both cases the function has a unique saddle point u_τ with $|u_\tau| < \mu$ and a steepest descent contour that is locally parallel to the imaginary axis.

After a change of variables $z = -ik$, Z can be written as a p-dimensional complex contour integral

$$Z = \frac{(-i\mu)^p}{(\pi\tau)^{\frac{p}{2}}\sqrt{\det(C)}} \int_{i\mathbb{R}^p} \frac{e^{\tau(z-w)^T C^{-1}(z-w)}}{\prod_{j=1}^p (\mu^2 - z_j^2)}\, dz. \tag{11}$$

Cauchy's theorem (Lang, 2002; Schneidemann, 2005) states that this integral remains invariant if the integration contours are deformed, as long as we remain in a domain where the integrand does not diverge (Appendix A.4). The analogue of Laplace's approximation for complex contour integrals, known as the stationary phase, steepest descent or saddle point approximation, then states that an integral of the form (11) can be approximated by a Gaussian integral along a steepest descent contour passing through the saddle point of the argument of the exponential function (Wong, 2001). Here, the function $(z - w)^T C^{-1}(z - w)$ has a saddle point at $z = w$. If $|w_j| < \mu$ for all j, the standard stationary phase approximation can be applied directly, but this only covers the uninteresting situation where the maximum-likelihood solution $\hat{x} = \operatorname{argmin}_x H(x) = 0$ (Appendix A.5). As soon as $|w_j| > \mu$ for at least one j, the standard argument breaks down, since to deform the integration contours from the imaginary axes to parallel contours passing through the saddle point $z_0 = w$, we would have to pass through a pole (divergence) of the function $\prod_j (\mu^2 - z_j^2)^{-1}$ (Figure 1). Motivated by similar, albeit one-dimensional, analyses in non-equilibrium physics (Lee et al., 2013), we instead consider a temperature-dependent function

$$H^*_\tau(z) = (z - w)^T C^{-1}(z - w) - \frac{1}{\tau}\sum_{j=1}^{p} \ln(\mu^2 - z_j^2), \tag{12}$$

which is well-defined on the domain $\mathcal{D} = \{z \in \mathbb{C}^p : |\Re z_j| < \mu,\ j = 1, \dots, p\}$, where $\Re$ denotes the real part of a complex number. This function has a unique saddle point in $\mathcal{D}$, regardless whether $|w_j| < \mu$ or not (Figure 1). Our main result is a steepest descent approximation of the partition function around this saddle point.

Theorem 1. *Let $C \in \mathbb{R}^{p\times p}$ be a positive definite matrix, $w \in \mathbb{R}^p$ and $\mu > 0$. Then the complex function H^*_τ defined in eq.* (12) *has a unique saddle point $\hat{u}_\tau$ that is real, $\hat{u}_\tau \in \mathcal{D} \cap \mathbb{R}^p$, and is a solution of the set* *of third order equations*

$$(\mu^2 - u_j^2)[C^{-1}(w-u)]_j - \frac{u_j}{\tau} = 0\,, \quad u \in \mathbb{R}^p,\ j \in \{1,\dots,p\}. \tag{13}$$

For $Q(z)$ a complex analytic function of $z \in \mathbb{C}^p$ diverging at most polynomially, i.e. $|Q(z)| \leq |z|^q$ for some $q \geq 0$, the generalized partition function

$$Z[Q] = \frac{\mu^p}{(\pi\tau)^{\frac{p}{2}}\sqrt{\det(C)}} \int_{\mathbb{R}^p} \frac{e^{-\tau(k-iw)^T C^{-1}(k-iw)} Q(-ik)}{\prod_{j=1}^p (k_j^2 + \mu^2)} dk.$$

can be analytically expressed as

$$Z[Q] = \Big(\frac{\mu}{\sqrt{\tau}}\Big)^p \frac{e^{\tau(w-\hat{u}_\tau)^T C^{-1}(w-\hat{u}_\tau)}}{\sqrt{\prod_{j=1}^p (\mu^2 + \hat{u}_{\tau,j}^2)\det(C + D_\tau)}} \exp\Big\{\frac{1}{4\tau^2}\Delta_\tau\Big\} e^{R_\tau(ik)} Q(\hat{u}_\tau + ik)\Big|_{k=0}, \tag{14}$$

where D_τ is a diagonal matrix with diagonal elements

$$(D_\tau)_{jj} = \frac{\tau(\mu^2 - \hat{u}_{\tau,j}^2)^2}{\mu^2 + \hat{u}_{\tau,j}^2}, \tag{15}$$

Δ_τ is the differential operator

$$\Delta_\tau = \sum_{i,j=1}^p \big[\tau D_\tau (C + D_\tau)^{-1} C\big]_{ij} \frac{\partial^2}{\partial k_i \partial k_j} \tag{16}$$

and

$$R_\tau(z) = \sum_{j=1}^p \sum_{m\geq 3} \frac{1}{m}\Big[\frac{1}{(\mu - \hat{u}_{\tau,j})^m} + \frac{(-1)^m}{(\mu + \hat{u}_{\tau,j})^m}\Big] z_j^m. \tag{17}$$

This results in an analytic approximation

$$Z[Q] \sim \Big(\frac{\mu}{\sqrt{\tau}}\Big)^p \frac{e^{\tau(w-\hat{u}_\tau)^T C^{-1}(w-\hat{u}_\tau)} Q(\hat{u}_\tau)}{\sqrt{\prod_{j=1}^p (\mu^2 + \hat{u}_{\tau,j}^2)\det(C + D_\tau)}} \tag{18}$$

The analytic expression in eq. (14) follows by changing the integration contours to pass through the saddle point $\hat{u}_\tau$, and using a Taylor expansion of $H_\tau^*(z)$ around the saddle point along the steepest descent contour; eq. (18) then results by taking the first-order term in the expansion of the differential operator exponential. However, because Δ_τ and R_τ depend on τ, it is not a priori evident that the higher-order terms in the exponential can be discarded. A detailed proof is given in Appendix B. The analytic approximation in eq. (18) can be simplified further by expanding $\hat{u}_\tau$ around its leading term, resulting in an expression that recognizably converges to the sparse maximum-likelihood solution (Appendix C). While eq. (18) is computationally more expensive to calculate than the corresponding expression in terms of the maximum-likelihood solution, it was found to be numerically more accurate (Section 3).

Various quantities derived from the posterior distribution can be expressed in terms of generalized partition functions. The most important of these are the expectation values of the regression coefficients, which, using elementary properties of the Fourier transform (Appendix A.6), can be expressed as

$$\mathbb{E}(x) = \frac{1}{Z}\int_{\mathbb{R}^p} x\, e^{-\tau H(x)} dx = \frac{Z\big[C^{-1}(w-z)\big]}{Z} \sim C^{-1}(w - \hat{u}_\tau).$$

The leading term,

$$\hat{x}_\tau \equiv C^{-1}(w - \hat{u}_\tau), \tag{19}$$

can be interpreted as an estimator for the regression coefficients in its own right, which interpolates smoothly (as a function of τ) between the ridge regression estimator $\hat{x}_{\text{ridge}} = C^{-1}w$ at

$\tau = 0$ and the maximum-likelihood elastic net estimator $\hat{x} = C^{-1}(w - \hat{u})$ at $\tau = \infty$, where $\hat{u} = \lim_{\tau\to\infty} \hat{u}_\tau$ is the solution of the constrained optimization problem

$$\hat{u} = \operatorname*{argmin}_{\{u\in\mathbb{R}^p\colon |u_j|\leq\mu,\forall j\}} (w-u)^T C^{-1}(w-u) \tag{20}$$

(see Michoel (2016) and Appendix C). Because $\hat{u}_\tau$ satisfies the convex optimization problem $\hat{u}_\tau = \operatorname{argmin}_{u\in\mathbb{R}^p} H^*_\tau(u)$, with H^*_τ defined in eq. (12) (see Appendix B.1), which corresponds to replacing the hard constraints $|u_j| \leq \mu$ in eq. (20) by a log-barrier function, Fenchel's convex duality theorem implies that $\hat{x}_\tau$ satisfies the convex optimization problem

$$\hat{x}_\tau = \operatorname*{argmin}_{x\in\mathbb{R}^p} x^T Cx - 2w^T x + 2\sum_{j=1}^{p} g_{\tau,\mu}(x_j), \tag{21}$$

where

$$g_{\tau,\mu}(t) = \frac{1}{\tau}\left(\sqrt{4\tau^2\mu^2t^2+1} - 1\right) + \frac{1}{\tau}\ln\left(\frac{\sqrt{1+4\tau^2\mu^2t^2}}{2\tau^2t^2}\right).$$

is a smooth approximation to the ℓ_1-penalty $\mu|t|$ for $t \in \mathbb{R}$ (Appendix D). With hindsight, this could be used to prove Theorem 1 using Laplace approximation techniques without the use of the Fourier transform.

Other quantities of interest are the marginal posterior distributions for subsets $I \subset \{1,\dots,p\}$ of regression coefficients, defined as

$$p(x_I) = \frac{1}{Z(C,w,\mu)} \int_{\mathbb{R}^{|I^c|}} e^{-\tau H(x|C,w,\mu)} dx_{I^c}$$

where $I^c = \{1,\dots,p\} \setminus I$ is the complement of I, $|I|$ denotes the size of a set I, and we have reintroduced temporarily the dependency on C, w and μ as in eq. (7). A simple calculation shows that the remaining integral is again a partition function of the same form, more precisely:

$$p(x_I) = e^{-\tau(x_I^T C_I x_I - 2w_I^T x_I + 2\mu\|x_I\|_1)} \frac{Z(C_{I^c}, w_{I^c} - x_I^T C_{I,I^c}, \mu)}{Z(C,w,\mu)}, \tag{22}$$

where the subscripts I and I^c indicate sub-vectors and sub-matrices on their respective coordinate sets. Hence the analytic approximation in eq. (14) can be used to approximate numerically each term in the partition function ratio and obtain an approximation to the marginal posterior distributions.

3. Numerical experiments

To test the accuracy of the stationary phase approximation, algorithms to solve the saddle point equations and compute the partition function and marginal posterior distribution, as well as an existing Gibbs sampler algorithm (Hans, 2011), were implemented in Matlab (see Appendix F for algorithm details; source code available from `https://github.com/tmichoel/bayonet/`). Results were evaluated for independent predictors (or equivalently, one predictor) and two commonly used data sets for testing lasso and elastic net algorithms: the "diabetes data", consisting of $p = 10$ baseline predictor variables for $n = 442$ diabetes patients, and a quantitative response measure of disease progression one year after baseline (Efron et al., 2004); and the "leukemia data", consisting of $p = 3571$ gene expression predictor variables for $n = 72$ leukemia samples, and a binary response variable indicating whether the sample is type 1 (ALL) or type 2 (AML) leukemia (Zou and Hastie, 2005) (see Appendix G for experimental details and data sources).

First the fundamental relation (cf. Appendix C)

$$\lim_{\tau\to\infty} -\frac{1}{\tau}\log Z = H_{\min} = \min_{x\in\mathbb{R}^p} H(x)$$

was tested. For independent predictors ($p = 1$), the partition function can be calculated analytically using the error function (Appendix E), and rapid convergence to $H_{\min}$ is observed (Figure 2a). After scaling by the number of predictors p, a similar rate of convergence is observed for the stationary phase approximation to the partition function for both the diabetes and leukemia data

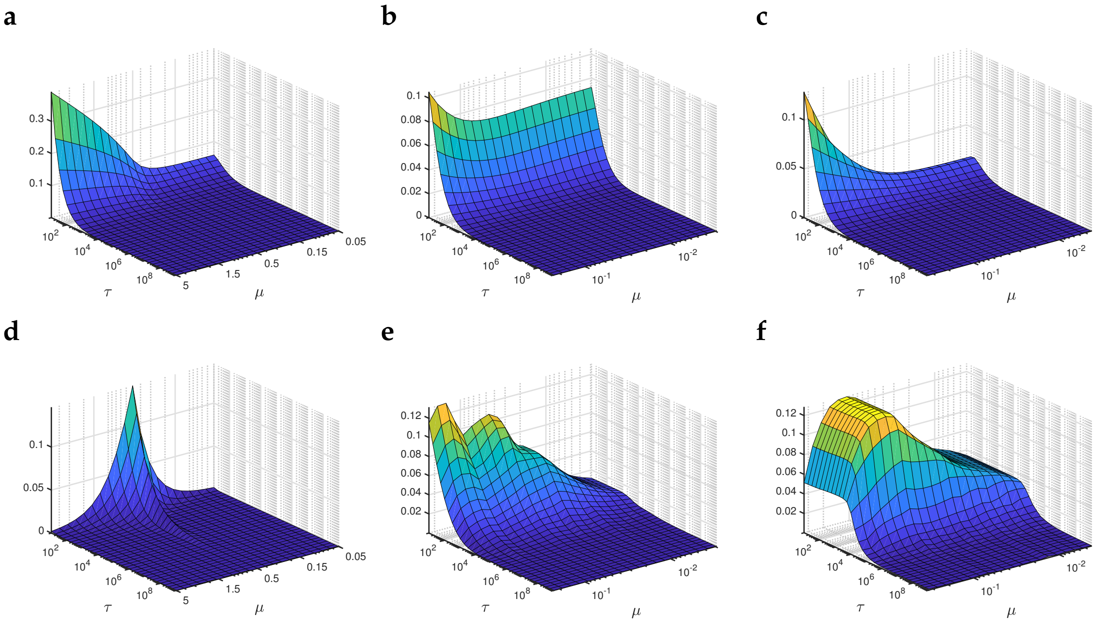

FIGURE 2. Convergence to the maximum-likelihood/minimum-energy solution. Top row: $(-\frac{1}{\tau}\log Z - H_{\min})/p$ vs. τ and μ for the exact partition function for independent predictors ($p = 1$, **a**), and for the stationary phase approximation to the partition function for the diabetes (**b**) and leukemia (**c**) data. Bottom row: $\|\hat{x}_\tau - \hat{x}\|_\infty$ for the exact expectation value for independent predictors (**d**), and using the stationary phase approximation for the diabetes (**e**) and leukemia (**f**) data. See Appendix G.2 and G.3 for experimental details.

(Figure 2b,c). However, convergence of the posterior expectation values $\hat{x}_\tau$ to the maximum-likelihood coefficients $\hat{x}$, as measured by the ℓ_∞-norm difference $\|\hat{x}_\tau - \hat{x}\|_\infty = \max_j |\hat{x}_{\tau,j} - \hat{x}_j|$ is noticeably slower, particularly in the $p \gg n$ setting of the leukemia data (Figure 2d–f).

Next, the accuracy of the stationary phase approximation at finite τ was determined by comparing the marginal distributions for single predictors [i.e. where I is a singleton in eq. (22)] to results obtained from Gibbs sampling. For simplicity, representative results are shown for specific hyper-parameter values (Appendix G.3). Application of the stationary phase approximation resulted in marginal posterior distributions which were indistinguishable from those obtained by Gibbs sampling (Figure 3). An approximation to eq. (22) of the form

$$p(x_I) \approx e^{-\tau(x_I^T C_I x_I - 2w_I^T x_I + 2\mu\|x_I\|_1)} e^{-\tau[H_{\min}(C_{I^c}, w_{I^c} - x_I^T C_{I,I^c}, \mu) - H_{\min}(C, w, \mu)]} \tag{23}$$

was also tested. However, while eq. (23) is indistinguishable from eq. (22) for predictors with zero effect size in the maximum-likelihood solution, it resulted in distributions that were squeezed towards zero for transition predictors, and often wildly inaccurate for non-zero predictors (Figure 3). This is because eq. (23) is easily seen to be maximized at $x_I = \hat{x}_I$, the global maximum-likelihood value, whereas the true marginal distributions need *not* be maximized at this value. Hence, accurate estimations of the marginal posterior distributions requires using the full stationary phase approximations [eq. (18)] to the partition functions in eq. (22). This does not contradict the rapid the convergence of the log-parition function to the minimum-energy value (Figure 2), because the latter is on a logarithmic scale, whereas the marginal distributions involve ratios of partition functions on an absolute scale.

The stationary phase approximation is particularly advantageous in prediction problems, where the response value $\tilde{y} \in \mathbb{R}$ for a newly measured predictor sample $\tilde{A} \in \mathbb{R}^{1\times p}$ is obtained using regression coefficients learned from training data (y, A). In Bayesian inference, $\tilde{y}$ is set to the expectation value of the posterior predictive distribution (Friedman et al., 2001), $\tilde{y} = \mathbb{E}(y) = \tilde{A}\mathbb{E}(x)$.

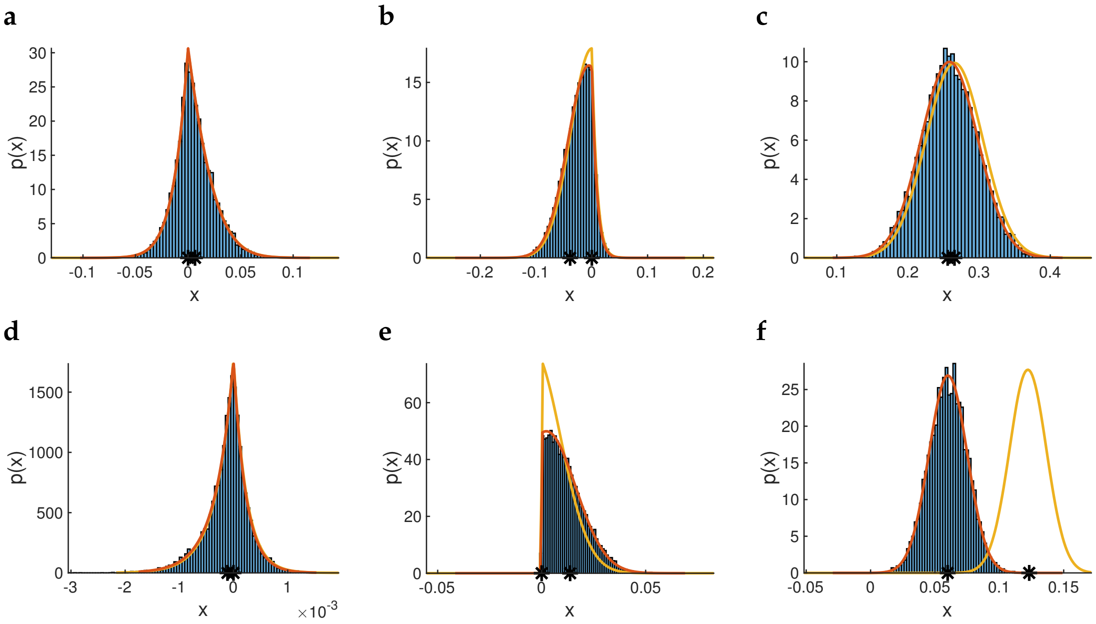

FIGURE 3. Marginal posterior distributions for the diabetes (**a–c**), and leukemia data (**d–f**). In red, stationary phase approximation for the marginal posterior distribution of selected predictors. In blue, Gibbs sampling histogram (10^4 samples). In yellow, maximum-likelihood-based approximation. Shown are the distributions for a zero, transition and non-zero maximum-likelihood predictor (from left to right). The stars on the x-axes indicate the location of the maximum-likelihood and posterior expectation value. See Appendix G.3 for experimental details.

Computation of the posterior expectation values $\mathbb{E}(x) = \hat{x}_\tau$ [eq. (19)] using the stationary phase approximation requires solving only one set of saddle point equations, and hence can be performed efficiently across a range of hyper-parameter values, in contrast to Gibbs sampling, where the full posterior needs to be sampled even if only expectation values are needed. To illustrate how this benefits large-scale applications of the Bayesian elastic net, the prediction performance of the Bayesian elastic net was compared to maximum-likelihood elastic net and ridge regression using gene expression and drug sensitivity data for 17 anticancer drugs in 474 human cancer cell lines from the Cancer Cell Line Encyclopedia (Barretina et al., 2012) (see Appendix G.4 for experimental details and data sources). Using 10-fold cross-validation across both μ and τ, the median correlation between predicted and true drug sensitivities was consistenly higher for the Bayesian elastic net than the maximum-likelihood elastic net and ridge regression ($\mu = 0$) (Figure 4a). While the difference in optimal performance between Bayesian and maximum-likelihood elastic net was not always large, Bayesian elastic net tended to be optimized at larger values of μ (i.e. at sparser maximum-likelihood solutions), and at these values the performance improvement over maximum-likelihood elastic net was particularly pronounced (Figure 4b and Supplementary Figures S1 and S2). As expected, τ acts as a tuning parameter that allows to smoothly vary from the maximum-likelihood solution at large τ (here, $\tau \sim 10^6$) to the solution with best cross-validation performance (here, $\tau \sim 10^3 - 10^4$) (Figure 4c and Supplementary Figures S1 and S2). The improved performance at sparsety-inducing values of μ suggests that the Bayesian elastic net is uniquely able to identify the dominant predictors for a given response (the non-zero maximum-likelihood coefficients), while still accounting for the cumulative contribution of predictors with small effects. Comparison with the unpenalized ($\mu = 0$) ridge regression coefficients shows that the Bayesian expectation values are strongly shrunk towards zero, except for the non-zero maximum-likelihood coefficients, which remain relatively unchanged (Figure 4d), resulting in a

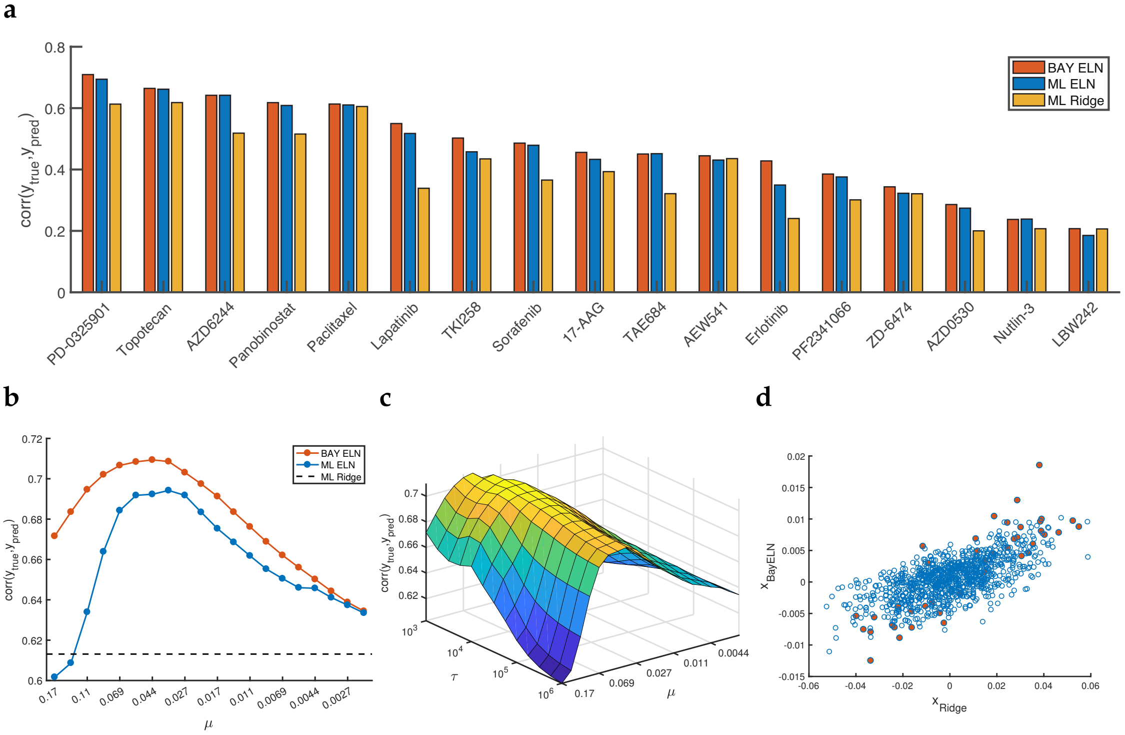

FIGURE 4. Predictive accuracy on the Cancer Cell Line Encyclopedia. **a.** Median correlation coefficient between predicted and true drug sensitivities over 10-fold cross-validation, using Bayesian posterior expectation values (red) and maximum-likelihood elastic net (blue) and ridge (yellow) regression values for the regression coefficients; $\lambda = 0.1$ was fixed, while μ and τ were optimized over 20, resp. 13 geometric values. **b.** Median 10-fold cross-validation value for the correlation coefficient between predicted and true sensitivities for the compound PD-0325901 vs. μ, for the Bayesian elastic net at optimal τ (red), maximum-likelihood elastic net (blue) and ridge regression (dashed). **c.** Median 10-fold cross-validation value for the correlation coefficient between predicted and true sensitivities for PD-0325901 for the Bayeslan elastic net vs. τ and μ. **d.** Scatter plot of expected regression coefficients in the Bayesian elastic net for PD-0325901 at $\mu = 0.068$ and optimal $\tau = 10^3$ vs. ridge regression coefficient estimates; coefficients with non-zero maximum-likelihood elastic net value at the same μ are indicated in red. See Appendix G.4 for experimental details.

double-exponential distribution for the regression coefficients. This contrasts with ridge regression, where regression coefficients are normally distributed leading to over-estimation of small effects, and maximum-likelihood elastic net, where small effects become identically zero and don't contribute to the predicted value.

4. CONCLUSIONS

The application of Bayesian methods to infer expected effect sizes and marginal posterior distributions in ℓ_1-penalized models has so far required the use of computationally expensive Gibbs sampling methods. Here it was shown that highly accurate inference in these models is actually possible using an analytic stationary phase approximation to the partition function integrals. This approximation exploits the fact that the Fourier transform of the non-differentiable double-exponential prior distribution is a well-behaved exponential of a log-barrier function, which is intimately related to the Legendre-Fenchel transform of the ℓ_1-penalty term. Thus, the Fourier transform is seen to play the same role for Bayesian inference problems as convex duality plays for maximum-likelihood approaches. For simplicity, we have focused on the linear regression model, where the invariance of multivariate normal distributions under the Fourier transform greatly facilitates the analytic derivations. However, it is clear that similar results are expected to hold for generalized linear and non-linear models (cf. Appendix A.2).

A limitation of the current approach may be that values of the hyper-parameters need to be specified in advance, whereas in complete hierarchical models, these would be subject to their own prior distributions. Incorporation of such priors will require careful attention to the interchange between taking the limit of and integrating over the inverse temperature parameter. In many practical situations though, researchers will perform maximum-likelihood inference and determine ℓ_1 and ℓ_2-penalty parameters by cross-validation or by specifying the level of sparsity. Setting the residual variance parameter to its maximum a-posteriori value then allows to evaluate the maximum-likelihood solution in the context of the posterior distribution of which it is the mode, while taking into account the amount of unexplained variance in the response, as has been suggested previously (Hans, 2011). Alternatively, in applications where the posterior expectation values of the regression coefficients are used instead of their maximum-likelihood values to predict unmeasured responses, the optimal inverse-temperature parameter can be determined by standard cross-validation on the training data.

No attempt was made to optimize the speed of the coordinate descent algorithm to solve the saddle point equations (Appendix F.1). However, comparison to the Gibbs sampling algorithm (Appendix F.5) shows that one cycle through all coordinates in the coordinate descent algorithm is approximately equivalent to one cycle in the Gibbs sampler, i.e. to adding one more sample. Empirically, it was found that the coordinate descent algorithm typically converges in 5-10 cycles starting from the maximum-likelihood solution, and 1-2 cycles when starting from a neighbouring solution in the estimation of marginal distributions (Appendix F.3). In contrast, Gibbs sampling typically requires 10^3-10^5 coordinate cycles to obtain stable distributions. Hence, in applications where only the posterior expectation values or the posterior distributions for a limited number of coordinates are sought, the computational advantage of the stationary phase approximation is vast. On the other hand, each evaluation of the marginal distribution functions requires the solution of a separate set of saddle point equations. Hence, computing these distributions for all predictors at a large number of points with the current algorithm could become equally expensive as Gibbs sampling. In practice, the need to evaluate all posterior distributions should occur rarely, because a more efficient maximum-likelihood-based approximation to the marginal posterior distributions was found to be accurate for the majority of predictors with zero maximum-likelihood effect sizes.

In summary, expressing intractable partition function integrals as complex-valued oscillatory integrals through the Fourier transform is a powerful approach for performing Bayesian inference in the lasso and elastic net regression models, and ℓ_1-penalized models more generally. Use of the stationary phase approximation to these integrals results in highly accurate estimates for the posterior expectation values and marginal distributions at a much reduced computational cost compared to Gibbs sampling.

Appendix A. Basic results in Fourier space

A.1. **Fourier transform conventions.** Fourier transforms are defined with different scaling conventions in different branches of science. Here, the symmetric version of the Fourier transform written in terms of angular frequencies is used: for f a function on $\mathbb{R}^p$, we define

$$\mathcal{F}[f](k) = \hat{f}(k) = \frac{1}{(2\pi)^{\frac{p}{2}}} \int_{\mathbb{R}^p} f(x) e^{-ik^T x} dx$$

and

$$f(x) = \mathcal{F}^{-1}\left[\mathcal{F}[f]\right](x) = \frac{1}{(2\pi)^{\frac{p}{2}}} \int_{\mathbb{R}^p} \hat{f}(k) e^{ik^T x} dk.$$

Parseval's identity states that for two functions f and g,

$$\int_{\mathbb{R}^p} \overline{f(x)} g(x) dx = \int_{\mathbb{R}^p} \overline{\hat{f}(k)} \hat{g}(k) dk,$$

where $\bar{\cdot}$ denotes complex conjugation. For more details, see (Hunter and Nachtergaele, 2001, Chapter 11).

A.2. **Relation between convex duality and the Fourier transform.** The motivation for using the Fourier transform to study Bayesian inference problems stems from the correspondence between the Fourier and Legendre-Fenchel transforms of convex functions. This correspondence is an example of so-called idempotent mathematics, and a survey of its history and applications can be found in Litvinov (2005), while a formal treatment along the lines below can be found in Fedoryuk (1971); a summary of analogous properties between the Legendre-Fenchel and Fourier transforms can also be found in Alonso and Forbes (1995). The basic argument is presented here, without any attempt at being complete or rigorous.

Let h be a convex function on $\mathbb{R}^p$ and assume it is sufficiently smooth for the statements below to hold without needing too much attention to detail. The Gibbs probability distribution for h at inverse temperature τ is defined as $p(x) = \frac{1}{Z} e^{-\tau h(x)}$, with $Z = \int_{\mathbb{R}^p} e^{-\tau h(x)} dx$ the partition function. Define for $z \in \mathbb{C}^p$

$$h^*_\tau(z) = \frac{1}{\tau} \ln \int_{\mathbb{R}^p} e^{-\tau[h(x) - z^T x]} dx.$$

By the Laplace approximation, it follows that for τ large and $u \in \mathbb{R}^p$, to leading order in τ,

$$h^*_\tau(u) \approx h^*(u) = \max_{x \in \mathbb{R}^p} [u^T x - h(x)], \tag{24}$$

the Legendre-Fenchel transform of h. The Fourier transform of $e^{-\tau h}$ is

$$\mathcal{F}\big[e^{-\tau h}\big](\tau k) = \frac{1}{(2\pi)^{\frac{p}{2}}} \int_{\mathbb{R}^p} e^{-\tau h(x)} e^{-i\tau k^T x} dx = \frac{e^{\tau h^*_\tau(-ik)}}{(2\pi)^{\frac{p}{2}}}. \tag{25}$$

Now assume that $h = f + g$ can be written as the sum of two convex functions f and g. It is instructive to think of $h(x)$ as minus a posterior log-likelihood function of regression coefficients x, with a natural decomposition in a part $f(x)$ coming from the data likelihood and a part $g(x)$ representing the prior distribution on x. We again assume that f and g are smooth.

The Parseval identity for Fourier transforms yields

$$\int_{\mathbb{R}^p} e^{-\tau[f(x)+g(x)]} dx = \int_{\mathbb{R}^p} \overline{\mathcal{F}\big[e^{-\tau f}\big](k)} \mathcal{F}\big[e^{-\tau g}\big](k) dk = \Big(\frac{\tau}{2\pi}\Big)^p \int_{\mathbb{R}^p} e^{\tau[f^*_\tau(ik) + g^*_\tau(-ik)]} dk,$$

where a change of variables $k \to \tau k$ was made. When τ is large, the Laplace approximation of the l.h.s. states that, to leading order in τ

$$\frac{1}{\tau} \ln \int_{\mathbb{R}^p} e^{-\tau[f(x)+g(x)]} dx \approx -\min_{x \in \mathbb{R}^p} \big[f(x) + g(x)\big] = \max_{x \in \mathbb{R}^p} \big[-f(x) - g(x)\big]. \tag{26}$$

The integral on the r.h.s. can be written as a complex contour integral

$$\int_{\mathbb{R}^p} e^{\tau[f^*_\tau(ik) + g^*_\tau(-ik)]} dk = \frac{1}{i^p} \int_{i\mathbb{R}^p} e^{\tau[f^*_\tau(z) + g^*_\tau(-z)]} dz,$$

where $i\mathbb{R}^p$ denotes a p-dimensional contour consisting of vertical contours running along the imaginary axis in each dimension. The steepest descent or saddle point approximation (Wong, 2001) requires that we deform the contour to run through the saddle point, i.e. a zero of the gradient function $\nabla[f^*_\tau(z) + g^*_\tau(-z)]$. Under fairly general conditions (see for instance Daniels (1954)), $f^*_\tau(z) + g^*_\tau(-z)$ will attain its maximum modulus at a real vector, and hence the new integration contour will take the form $z = \hat{u}_\tau + ik$ where $\hat{u}_\tau = \operatorname{argmin}_{u \in \mathbb{R}^p} [f^*_\tau(u) + g^*_\tau(-u)]$ and

$k \in \mathbb{R}^p$. Note that in the limit $\tau \to \infty$, $\hat{u}_\tau \to \hat{u} = \operatorname{argmin}_{u\in\mathbb{R}^p}[f^*(u) + g^*(-u)]$. The stationary phase approximation yields, again to leading order in τ

$$\begin{aligned}\frac{1}{\tau}\ln\int_{\mathbb{R}^p} e^{\tau[f^*_\tau(ik)+g^*_\tau(-ik)]}dk &= \frac{1}{\tau}\ln\int_{\mathbb{R}^p} e^{\tau[f^*_\tau(\hat{u}_\tau+ik)+g^*_\tau(-\hat{u}_\tau-ik)]}dk \\ &\approx \min_{u\in\mathbb{R}^p}\left[f^*_\tau(u)+g^*_\tau(-u)\right] \approx \min_{u\in\mathbb{R}^p}\left[f^*(u)+g^*(-u)\right] \end{aligned} \quad (27)$$

Combining eqs. (26) and (27), we recover Fenchel's well-known duality theorem

$$\max_{x\in\mathbb{R}^p}\left[-f(x)-g(x)\right] = \min_{u\in\mathbb{R}^p}\left[f^*(u)+g^*(-u)\right].$$

In summary, there is an equivalence between convex duality for log-likelihood functions and switching from coordinate to frequency space using the Fourier transform for Gibbs probability distributions, which becomes an exact mapping in the limit of large inverse temperature. As shown in this paper, this remains true even when f or g are not necessarily smooth (e.g. if $g(x) = \|x\|_1$ is the ℓ_1-norm).

A.3. **The Fourier transform of the multivariate normal and Laplace distributions.** To derive eq. (10), observe that $f(x)$ is a Gaussian and its Fourier transform is again a Gaussian:

$$\overline{\mathcal{F}(e^{-2\tau f})} = \frac{1}{(2\pi)^{\frac{p}{2}}}\int_{\mathbb{R}^p} e^{-2\tau f(x)}e^{ik^Tx}dx = \frac{1}{\sqrt{(2\tau)^p\det(C)}}\exp\left\{-\frac{1}{4\tau}(k-2i\tau w)^TC^{-1}(k-2i\tau w)\right\}. \quad (28)$$

To calculate the Fourier transform of $e^{-\tau g}$, note that in one dimension

$$\int_{\mathbb{R}} e^{-\gamma|x|}e^{-ikx}dx = \frac{2\gamma}{k^2+\gamma^2},$$

and hence

$$\mathcal{F}(e^{-2\tau g})(k) = \frac{1}{(2\pi)^{\frac{p}{2}}}\prod_{j=1}^p\frac{4\mu\tau}{k_j^2+4\tau^2\mu^2}.$$

After making the change of variables $k'_j = \frac{1}{2\tau}k_j$, eq. (10) is obtained.

A.4. **Cauchy's theorem in coordinate space.** Cauchy's theorem (Lang, 2002; Schneidemann, 2005) states that we can freely deform the integration contours in the integral in eq. (11) as long as we remain within a holomorphic domain of the integrand, or simply put, a domain where the integrand does not diverge. Consider as a simple example the deformation of the integration contours from $z_j \in i\mathbb{R}$ in eq. (11) to $z_j \in w'_j + i\mathbb{R}$, where $|w'_j| < \mu$ for all j. We obtain

$$\begin{aligned} Z &= \frac{(-i\mu)^p}{(\pi\tau)^{\frac{p}{2}}\sqrt{\det(C)}}\int_{w'_1-i\infty}^{w'_1+i\infty}\cdots\int_{w'_p-i\infty}^{w'_p+i\infty} e^{\tau(z-w)^TC^{-1}(z-w)}\prod_{j=1}^p\frac{1}{\mu^2-z_j^2}\,dz_1\dots dz_p \\ &= \frac{\mu^p}{(\pi\tau)^{\frac{p}{2}}\sqrt{\det(C)}}\int_{\mathbb{R}^p} e^{-\tau(w'-w+ik)^TC^{-1}(w'-w+ik)}\prod_{j=1}^p\frac{1}{\mu^2-(w'_j+ik_j)^2}\,dk, \end{aligned}$$

where we parameterized $z_j = w'_j + ik_j$. Using the inverse Fourier transform, and reversing the results from Section 2 and Appendix A.3, we can write this expression as

$$Z = \int_{\mathbb{R}^p} e^{-2\tau\tilde{f}(x)}e^{-2\tau\tilde{g}(x)},$$

where

$$\tilde{f}(x) = \frac{1}{2}x^TCx - (w-w')^Tx \quad (29)$$

$$\tilde{g}(x) = \sum_{j=1}^p(\mu|x_j| - w'_jx_j). \quad (30)$$

Comparison with eqs. (8)–(9) shows that the freedom to deform the integration contour in Fourier space corresponds to an equivalent freedom to split $e^{-\tau H(x)}$ into a product of two functions. Clearly eq. (30) only defines an integrable function $e^{-2\tau\tilde{g}}$ if $|w'_j| < \mu$ for all j, which of course corresponds to the limitation imposed by Cauchy's theorem that the deformation of the integration contours cannot extend beyond the domain where the function $\prod_j(\mu^2 - z_j^2)^{-1}$ remains finite.

A.5. **Stationary phase approximation in the zero-effect case.** Assume that $|w_j| < \mu$ for all j. It then follows immediately that the maximum-likelihood or minimum-energy solution $\hat{x} = \text{argmin}_x H(x) = 0$. As above, we can deform the integration contours in (11) into steepest descent contours passing through the saddle point $z_0 = w$ of the function $h(z) = (z-w)^T C^{-1}(z-w)$ (cf. Figure 1a). We obtain

$$\begin{aligned} Z &= \frac{(-i\mu)^p}{(\pi\tau)^{\frac{p}{2}}\sqrt{\det(C)}} \int_{w_1 - i\infty}^{w_1+i\infty} \dots \int_{w_p-i\infty}^{w_p+i\infty} e^{\tau(z-w)^T C^{-1}(z-w)} \prod_{j=1}^p \frac{1}{\mu^2 - z_j^2}\, dz_1 \dots dz_p \\ &= \frac{\mu^p}{(\pi\tau)^{\frac{p}{2}}\sqrt{\det(C)}} \int_{\mathbb{R}^p} e^{-\tau k^T C^{-1} k} \prod_{j=1}^p \frac{1}{\mu^2 - (w_j + ik_j)^2}\, dk, \end{aligned} \tag{31}$$

where we parameterized $z_j = w_j + ik_j$. This integral can be written as a series expansion using the following standard result, included here for completeness.

Lemma 1. *Let $C \in \mathbb{R}^p \times \mathbb{R}^p$ be a positive definite matrix and let Δ_C be the differential operator*

$$\Delta_C = \sum_{i,j=1}^p C_{ij} \frac{\partial^2}{\partial k_i \partial k_j}.$$

Then

$$\frac{1}{\pi^{\frac{p}{2}}\sqrt{\det(C)}} \int_{\mathbb{R}^p} e^{-k^T C^{-1} k} \hat{f}(k) dk = \left(e^{\frac{1}{4}\Delta_C} \hat{f}\right)(0).$$

Proof. First note that

$$\Delta_C e^{-ik^T x} = -\sum_{ij} C_{ij} x_i x_j e^{-ik^T x} = -(x^T C x)\, e^{-ik^T x}, \tag{32}$$

i.e. $e^{ik^T x}$ is an 'eigenfunction' of Δ_C with eigenvalue $-(x^T C x)$, and hence

$$e^{\frac{1}{4}\Delta_C} e^{-ik^T x} = e^{-\frac{1}{4}x^T C x} e^{-ik^T x}.$$

Using the (inverse) Fourier transform, we can define

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}}} \int_{\mathbb{R}^p} \hat{f}(k) e^{ik^T x} dk,$$

and write

$$\hat{f}(k) = \frac{1}{(2\pi)^{\frac{p}{2}}} \int_{\mathbb{R}^p} f(x) e^{-ik^T x} dx.$$

Hence

$$\left(e^{\frac{1}{4}\Delta_C} \hat{f}\right)(k) = \frac{1}{(2\pi)^{\frac{p}{2}}} \int_{\mathbb{R}^p} f(x) e^{\frac{1}{4}\Delta_C} e^{ik^T x} dx = \frac{1}{(2\pi)^{\frac{p}{2}}} \int_{\mathbb{R}^p} f(x) e^{-\frac{1}{4}x^T C x} e^{-ik^T x} dx.$$

Using Parseval's identity and the formula for the Fourier transform of a Gaussian [eq. (28)], we obtain

$$\left(e^{\frac{1}{4}\Delta_C} \hat{f}\right)(0) = \frac{1}{(2\pi)^{\frac{p}{2}}} \int_{\mathbb{R}^p} f(x) e^{-\frac{1}{4}x^T C x} dx = \frac{1}{\pi^{\frac{p}{2}}\sqrt{\det(C)}} \int_{\mathbb{R}^p} \hat{f}(k) e^{-k^T C^{-1} k} dk$$

□

In the derivation above, we have tacitly assumed that the inverse Fourier transform f of $\hat{f}$ exists. However, the result remains true even if f is only a distribution, i.e. $\hat{f}$ need not be integrable. For a more detailed discussion, see (Hunter and Nachtergaele, 2001, Chapter 11, Section 11.9).

Applying Lemma 1 to eq. (31), it follows that

$$Z = \Bigl(\frac{\mu}{\tau}\Bigr)^p e^{\frac{1}{4\tau}\Delta_C} \prod_{j=1}^p \frac{1}{\mu^2 - (w_j + ik_j)^2}\Big|_{k=0} = \Bigl(\frac{\mu}{\tau}\Bigr)^p \Bigl[\prod_{j=1}^p \frac{1}{\mu^2 - w_j^2} + \mathcal{O}\bigl(\frac{1}{\tau}\bigr)\Bigr],$$

with Δ_C as defined in eq. (32). It follows that the effect size expectation values are, to first order in τ^{-1},

$$\mathbb{E}(x_j) = \frac{1}{2\tau}\frac{\partial \log Z}{\partial w_j} \sim \frac{1}{\tau}\frac{w_j}{\mu^2 - w_j^2},$$

which indeed converge to the minimum-energy solution $\hat{x} = 0$.

A.6. **Generalized partition functions for the expected effects.** Using elementary properties of the Fourier transform, it follows that

$$\mathcal{F}\bigl[x_j e^{-2\tau f(x)}\bigr](k) = i\frac{\partial \mathcal{F}\bigl[e^{-2\tau f(x)}\bigr](k)}{\partial k_j}, \tag{33}$$

with f defined in eq. (8), and hence, repeating the calculations leading up to eq. (10), we find

$$\mathbb{E}(x_j) = \frac{\int_{\mathbb{R}^p} x_j e^{-\tau H(x)} dx}{\int_{\mathbb{R}^p} e^{-\tau H(x)} dx} = \frac{Z\Bigl[\bigl(C^{-1}(w-z)\bigr)_j\Bigr]}{Z} \sim \bigl[C^{-1}(w - \hat{u}_\tau)\bigr]_j. \tag{34}$$

Note that eq. (33) can also be applied to the Laplacian part $e^{-2\tau g(x)}$, with g defined in eq. (9). This results in

$$\mathbb{E}(x_j) = \frac{Z\Bigl[\frac{z_j}{\tau(\mu^2 - z_j^2)}\Bigr]}{Z} \sim \frac{\hat{u}_{\tau,j}}{\tau(\mu^2 - \hat{u}_{\tau,j}^2)}. \tag{35}$$

By the saddle point equations, eq. (13), eqs. (34) and (35) are identical. As a rule of thumb, 'tricks' such as eq. (33) to express properties of the posterior distribution as generalized partition functions lead to accurate approximations if the final result does not depend on whether the trick was applied to the Gaussian or Laplacian part of the Gibbs factor. For higher-order moments of the posterior distribution, this means that the leading term of the stationary phase approximation alone is not sufficient.

Appendix B. Proof of Theorem 1

B.1. **Saddle-point equations.** Consider the function H^*_τ defined in eq. (12),

$$H^*_\tau(z) = (z-w)^T C^{-1}(z-w) - \frac{1}{\tau}\sum_{j=1}^p \ln(\mu^2 - z_j^2),$$

with z restricted to the domain $\mathcal{D} = \{z \in \mathbb{C}^p \colon |\Re z_j| < \mu,\ j = 1,\dots,p\}$. Writing $z = u + iv$, where u and v are the real and imaginary parts of z, respectively, we obtain

$$\Re H^*_\tau(z) = (u-w)^T C^{-1}(u-w) - v^T C^{-1} v - \frac{1}{2\tau}\sum_{j=1}^p \Bigl\{\ln\bigl[(\mu+u_j)^2 + v_j^2)\bigr] + \ln\bigl[(\mu-u_j)^2 + v_j^2)\bigr]\Bigr\}$$

$$\Im H^*_\tau(z) = 2(u-w)^T C^{-1} v - \frac{1}{\tau}\sum_{j=1}^p \Bigl\{\arctan\bigl(\frac{v_j}{\mu+u_j}\bigr) + \arctan\bigl(\frac{v_j}{\mu-u_j}\bigr)\Bigr\},$$

where $\Re c$ and $\Im c$ denote the real and imaginary parts of a complex number c, respectively.

By the Cauchy-Riemann equations $z = u + iv$ is a saddle point of H^*_τ if and only if it satisfies the equations

$$\frac{\partial \Re H^*_\tau}{\partial u_j} = 2[C^{-1}(u-w)]_j - \frac{1}{\tau}\Big\{\frac{\mu+u_j}{(\mu+u_j)^2+v_j^2} - \frac{\mu-u_j}{(\mu-u_j)^2+v_j^2}\Big\} = 0$$
$$\frac{\partial \Re H^*_\tau}{\partial v_j} = -2[C^{-1}v]_j - \frac{1}{\tau}\Big\{\frac{v_j}{(\mu+u_j)^2+v_j^2} + \frac{v_j}{(\mu-u_j)^2+v_j^2}\Big\} = 0$$

The second set of equations is solved by $v = 0$, and because $\Re H^*_\tau(u+iv) < \Re H^*_\tau(u)$ for all u and $v \neq 0$, it follows that $v = 0$ is the saddle point solution. Plugging this into the first set of equations gives

$$[C^{-1}(u-w)]_j + \frac{u_j}{\tau(\mu^2-u_j^2)} = 0, \tag{36}$$

which is equivalent to eq. (13).

B.2. **Analytic expression for the partition function.** Next, consider the complex integral

$$\mathcal{I} = (-i)^p \int_{-i\infty}^{i\infty} \cdots \int_{-i\infty}^{i\infty} e^{\tau H^*_\tau(z)} Q(z) dz_1 \dots dz_p,$$

i.e. $\mathcal{I}$ is the generalized partition function upto a constant multiplicative factor. By Cauchy's theorem we can freely deform the integration contours to a set of vertical contours running parallel to the imaginary axis and passing through the saddle point, i.e. integrate over $z = \hat{u}_\tau + ik$, where $\hat{u}_\tau$ is the saddle point solution and $k \in \mathbb{R}^p$. Changing the integration variable back from complex z to real k, we find

$$\mathcal{I} = e^{\tau(w-\hat{u}_\tau)C^{-1}(w-\hat{u}_\tau)} \int_{\mathbb{R}^p} e^{-\tau F(k)} Q(\hat{u}_\tau + ik) dk$$

where

$$F(k) = k^T C^{-1} k - 2ik^T C^{-1}(\hat{u}_\tau - w) + \frac{1}{\tau}\sum_{j=1}^p \ln(\mu - \hat{u}_{\tau,j} - ik_j) + \frac{1}{\tau}\sum_{j=1}^p \ln(\mu + \hat{u}_{\tau,j} + ik_j).$$

First we show that the main contribution to the integral in $\mathcal{I}$ comes from a small region around $k = 0$. This is true in fact for *any* set of vertical contours, not only those passing through the saddle point, and follows from standard arguments for the Laplace approximation (Wong, 2001).

Lemma 2. *Let $u \in \mathbb{R}^p$ with $|u_j| < \mu$ for all j, $\tilde{Q}$ a complex analytic function on $\mathbb{R}^p$ with $|\tilde{Q}(z)| \leq |z|^q$ from some $q \geq 0$, D_0 a compact subdomain of $\mathbb{R}^p$ containing $k = 0$, and $\tau_0 > 0$. Then for $\tau > \tau_0$,*

$$\left| \int_{\mathbb{R}^p \setminus D_0} e^{-\tau(k^T C^{-1} k - 2ik^T C^{-1}(u-w))} \frac{\tilde{Q}(ik)}{\prod_{j=1}^p (\mu^2 - (u_j + ik_j)^2)} dk \right| \leq K e^{-(\tau-\tau_0)c},$$

where

$$c = \min_{k \in \mathbb{R}^p \setminus D_0} k^T C^{-1} k > 0$$
$$K = \int_{\mathbb{R}^p} \frac{e^{-\tau_0 k^T C^{-1} k} |Q(ik)|}{\prod_{j=1}^p \big[(\mu^2 - u_j^2 + ik_j^2)^2 + 4u_j^2 k_j^2\big]^{\frac{1}{2}}} dk < \infty$$

Proof.

$$\begin{aligned}
&\left|\int_{\mathbb{R}^p\setminus D_0} e^{-\tau(k^TC^{-1}k-2ik^TC^{-1}(u-w))}\frac{\tilde{Q}(ik)}{\prod_{j=1}^p(\mu^2-(u_j+ik_j)^2)}dk\right|\\
&\quad\leq \int_{\mathbb{R}^p\setminus D_0}\frac{e^{-\tau k^TC^{-1}k}\left|Q(ik)\right|}{\prod_{j=1}^p\left[(\mu^2-u_j^2+k_j^2)^2+4u_j^2k_j^2\right]^{\frac{1}{2}}}dk\\
&\quad\leq e^{-(\tau-\tau_0)c}\int_{\mathbb{R}^p\setminus D_0}\frac{e^{-\tau_0 k^TC^{-1}k}\left|Q(ik)\right|}{\prod_{j=1}^p\left[(\mu^2-u_j^2+k_j^2)^2+4u_j^2k_j^2\right]^{\frac{1}{2}}}dk\\
&\quad\leq e^{-(\tau-\tau_0)c}\int_{\mathbb{R}^p}\frac{e^{-\tau_0 k^TC^{-1}k}\left|Q(ik)\right|}{\prod_{j=1}^p\left[(\mu^2-u_j^2+k_j^2)^2+4u_j^2k_j^2\right]^{\frac{1}{2}}}dk = Ke^{-(\tau-\tau_0)c}.
\end{aligned}$$

That $c > 0$ and $K < \infty$ follows immediately from the assumptions of the Lemma. □

Lemma 2 implies that we can restrict the integral in $\mathcal{I}$ to a small domain around $k = 0$, or equivalently, that we may henceforth assume that $Q(\hat{u}_\tau + ik)$ has compact support.

Next we compute the Taylor series for F around $k = 0$. First note that the n^{th} derivative of $f_j^{\pm}(k_j) = \ln(\mu \pm \hat{u}_{\tau,j} \pm ik_j)$ evaluated at $k_j = 0$ is given by

$$(f_j^{\pm})^{(n)}(0) = -\frac{(\mp i)^n(n-1)!}{(\mu\pm\hat{u}_{\tau,j})^n}.$$

By the saddle point equations (36)

$$\frac{1}{\tau}\sum_{j=1}^p f_j^{+'}(0)k_j + \frac{1}{\tau}\sum_{j=1}^p f_j^{-'}(0)k_j = \frac{i}{\tau}\sum_{j=1}^p\frac{k_j}{\mu+\hat{u}_{\tau,j}} - \frac{i}{\tau}\sum_{j=1}^p\frac{k_j}{\mu-\hat{u}_{\tau,j}} = 2ik^TC^{-1}(\hat{u}_{\tau,j}-w).$$

Hence the linear terms cancel and we obtain

$$\begin{aligned}
F(k) &= \frac{1}{\tau}\sum_{j=1}^p\left[\ln(\mu+\hat{u}_{\tau,j})+\ln(\mu-\hat{u}_{\tau,j})\right] + k^TC^{-1}k + \frac{1}{\tau}\sum_{j=1}^p\frac{\mu^2+\hat{u}_{\tau,j}^2}{(\mu^2-\hat{u}_{\tau,j}^2)^2}k_j^2\\
&\quad -\frac{1}{\tau}\sum_{j=1}^p\sum_{n\geq 3}\frac{1}{n}\Big[\frac{1}{(\mu-\hat{u}_{\tau,j})^n}+\frac{(-1)^n}{(\mu+\hat{u}_{\tau,j})^n}\Big](ik_j)^n\\
&= \frac{1}{\tau}\sum_{j=1}^p\ln(\mu^2-\hat{u}_{\tau,j}^2) + k^T(C^{-1}+D_\tau^{-1})k - \frac{1}{\tau}R_\tau(ik),
\end{aligned}$$

with D_τ the diagonal matrix defined in eq. (15) and R_τ the function defined in eq. (17). Hence

$$\mathcal{I} = e^{\tau(w-\hat{u}_\tau)C^{-1}(w-\hat{u}_\tau)}\prod_{j=1}^p\frac{1}{\mu^2-\hat{u}_{\tau,j}^2}\int_{\mathbb{R}^p}e^{-\tau k^T(C^{-1}+D_\tau^{-1})k}e^{R_\tau(ik)}Q(\hat{u}_\tau+ik)dk.$$

Application of Lemma 1 results in

$$\begin{aligned}
&\int_{\mathbb{R}^p} e^{-\tau k^T(C^{-1}+D_\tau^{-1})k} e^{R_\tau(ik)} Q(\hat{u}_\tau + ik) dk \\
&\quad = \frac{(2\pi)^{\frac{p}{2}}}{(2\tau)^{\frac{p}{2}}\sqrt{\det(C^{-1}+D_\tau^{-1})}} \exp\Big\{\frac{1}{4\tau^2}\Delta_\tau\Big\} e^{R_\tau(ik)} Q(\hat{u}_\tau + ik)\Big|_{k=0} \\
&\quad = \Big(\frac{\pi}{\tau}\Big)^{\frac{p}{2}} \Big(\frac{\det(D_\tau)\det(C)}{\det(C+D_\tau)}\Big)^{\frac{1}{2}} \exp\Big\{\frac{1}{4\tau^2}\Delta_\tau\Big\} e^{R_\tau(ik)} Q(\hat{u}_\tau + ik)\Big|_{k=0} \\
&\quad = \pi^{\frac{p}{2}} \frac{\prod_j(\mu^2 - \hat{u}_{\tau,j}^2)}{\prod_j(\mu^2 + \hat{u}_{\tau,j}^2)^{\frac{1}{2}}} \Big(\frac{\det(C)}{\det(C+D_\tau)}\Big)^{\frac{1}{2}} \exp\Big\{\frac{1}{4\tau^2}\Delta_\tau\Big\} e^{R_\tau(ik)} Q(\hat{u}_\tau + ik)\Big|_{k=0},
\end{aligned}$$

where we used the equality

$$C^{-1} + D_\tau^{-1} = C^{-1}(C + D_\tau)D_\tau^{-1},$$

and Δ_τ is the differential operator defined in eq. (16). Hence

$$\begin{aligned}
Z[Q] &= \frac{\mu^p}{(\pi\tau)^{\frac{p}{2}}\sqrt{\det(C)}} \mathcal{I} \\
&= \Big(\frac{\mu}{\sqrt{\tau}}\Big)^p \frac{1}{\prod_j(\mu^2 + \hat{u}_{\tau,j}^2)^{\frac{1}{2}}} \frac{e^{\tau(w-\hat{u}_\tau)C^{-1}(w-\hat{u}_\tau)}}{\sqrt{\det(C+D_\tau)}} \exp\Big\{\frac{1}{4\tau^2}\Delta_\tau\Big\} e^{R_\tau(ik)} Q(\hat{u}_\tau + ik)\Big|_{k=0}.
\end{aligned}$$

Note that application of Lemma 1 requires the existence of the inverse Fourier transform of $e^{R_\tau(ik)}Q(\hat{u}_\tau + ik)$, at least a a tempered distribution. This is the case because by Lemma 2, we may assume that Q has compact support.

B.3. **Asymptotic properties of the saddle point.** Let $\hat{u} = \lim_{\tau\to\infty} \hat{u}_\tau$. By continuity, $\hat{u}$ is a solution to the set of equations

$$(u_j - \mu)(u_j + \mu)\big[C^{-1}(u - w)\big]_j = 0 \tag{37}$$

subject to the constraints $|u_j| \leq \mu$. Denote by $I \subseteq \{1, \dots, p\}$ the subset of indices j for which $\big[C^{-1}(\hat{u} - w)\big]_j \neq 0$. To facilitate notation, for $v \in \mathbb{R}^p$ a vector, denote by $v_I \in \mathbb{R}^{|I|}$ the sub-vector corresponding to the indices in I. Likewise denote by $C_I \in \mathbb{R}^{|I|\times|I|}$ the corresponding sub-matrix and by C_I^{-1} the inverse of C_I, i.e. $C_I^{-1} = (C_I)^{-1} \neq (C^{-1})_I$. Temporarily denoting $B = C^{-1}$, we can then rewrite the equations for $\hat{u}$ as

$$\begin{aligned}
\hat{u}_I &= \pm\mu \\
\big[C^{-1}(\hat{u} - w)\big]_{I^c} &= [B(\hat{u} - w)]_{I^c} = B_{I^c}(\hat{u}_{I^c} - w_{I^c}) + B_{I^c I}(\hat{u}_I - w_I) = 0,
\end{aligned}$$

or, using standard results for the inverse of a partitioned matrix (Horn and Johnson, 1985),

$$\hat{u}_{I^c} = w_{I^c} + B_{I^c}^{-1} B_{I^c I}(w_I - \hat{u}_I) = w_{I^c} - C_{I^c I} C_I^{-1}(w_I - \hat{u}_I).$$

Finally, define $\hat{x} = C^{-1}(w - \hat{u})$, and note that

$$\begin{aligned}
\hat{x}_I &= [B(w - \hat{u})]_I = B_I(w_I - \hat{u}_I) + B_{II^c}(w_{I^c} - \hat{u}_{I^c}) = (B_I - B_{II^c}B_{I^c}^{-1}B_{I^c,I})(w_I - \hat{u}_I) \\
&= C_I^{-1}(w_I - \hat{u}_I) \neq 0 \qquad (38) \\
\hat{x}_{I^c} &= 0. \qquad (39)
\end{aligned}$$

As we will see below, $\hat{x} = \operatorname{argmin}_{x\in\mathbb{R}^p} H(x)$ is the maximum-likelihood lasso or elastic net solution (cf. Appendix C), and hence the set I corresponds to the set of non-zero coordinates in this solution. Note that it is possible to have $\hat{u}_j = \pm\mu$ for $j \in I^c$ (i.e. $\hat{x}_j = 0$). This happens when μ is exactly at the transition value where j goes from not being included to being included in the ML solution. We will denote the subsets of I^c of transition and non-transition coordinates as I_t^c and I_{nt}^c, respectively. We then have the following lemma:

Lemma 3. *In the limit $\tau \to \infty$, we have*

$$\tau(\mu^2 - \hat{u}_{\tau,j}^2)^2 = \begin{cases} \mathcal{O}(\tau^{-1}) & j \in I \\ \mathcal{O}\big[(\tau \hat{x}_{\tau,j}^2)^{-1}\big] & j \in I_t^c \\ \mathcal{O}(\tau) & j \in I_{nt}^c \end{cases} \tag{40}$$

Proof. From the saddle point equations, we have

$$\tau(\mu^2 - \hat{u}_{\tau,j}^2)^2 = \frac{1}{\tau}\Big(\frac{\hat{u}_{\tau,j}}{\hat{x}_{\tau,j}}\Big)^2.$$

If $j \in I$, $\hat{x}_{\tau,j} \to \hat{x}_j \neq 0$ and $\hat{u}_{\tau,j} \to \hat{u}_j = \pm\mu$, and hence $\tau(\mu^2 - \hat{u}_{\tau,j}^2)^2 = \mathcal{O}(\tau^{-1})$. If $j \in I_{nt}^c$, $\mu^2 - \hat{u}_{\tau,j}^2 \to \mu^2 - \hat{u}_j^2 > 0$, and hence $\tau(\mu^2 - \hat{u}_{\tau,j}^2)^2 = \mathcal{O}(\tau)$. If $j \in I_t^c$, $\hat{x}_{\tau,j} \to 0$ and $\hat{u}_{\tau,j} \to \hat{u}_j = \pm\mu$, and hence $\tau(\mu^2 - \hat{u}_{\tau,j}^2)^2 = \mathcal{O}\big[(\tau\hat{x}_{\tau,j}^2)^{-1}\big]$. □

B.4. **Asymptotic properties of the differential operator matrix.** Let

$$E_\tau = \tau D_\tau (C + D_\tau)^{-1} C = \frac{\tau}{2}\big[D_\tau (C + D_\tau)^{-1} C + C(C + D_\tau)^{-1} D_\tau\big], \tag{41}$$

where the second equality is simply to make the symmetry of E_τ explicit. We have the following result:

Proposition 1. *Using the block matrix notation introduced above, and assuming $I_t^c = \emptyset$, the leading term of E_τ in the limit $\tau \to \infty$ can be written as*

$$E_\tau \sim \tau \begin{pmatrix} D_{\tau,I} & \frac{1}{2} D_{\tau,I} C_I^{-1} C_{II^c} \\ \frac{1}{2} D_{\tau,I} C_I^{-1} C_{II^c} & (C^{-1})_{I^c} \end{pmatrix}, \tag{42}$$

where I is again the set of non-zero coordinates in the maximum-likelihood solution.

Proof. Again using standard properties for the inverse of a partitioned matrix (Horn and Johnson, 1985), and the fact that D_τ is a diagonal matrix, we have for any index subset J

$$\big[(C + D_\tau)^{-1}\big]_J = \big[C_J + D_{\tau,J} - C_{J,J^c}(C_{J^c} + D_{\tau,J^c})^{-1} C_{J^c,J}\big]^{-1} \tag{43}$$

$$\big[(C + D_\tau)^{-1}\big]_{J,J^c} = -(C_J + D_{\tau,J})^{-1} C_{J^c,J} \big[(C + D_\tau)^{-1}\big]_{J^c} \tag{44}$$

By Lemma 3, in the limit $\tau \to \infty$, D_τ vanishes on I and diverges on I^c. Hence

$$(C_I + D_{\tau,I})^{-1} \sim C_I^{-1} \tag{45}$$

$$(C_{I^c} + D_{\tau,I^c})^{-1} \sim D_{\tau,I^c}^{-1} \tag{46}$$

Plugging these in eqs. (43) and (44), and using the fact that $C_{I,I^c} D_{\tau,I^c}^{-1} C_{I^c,I}$ is vanishingly small compared to C_I, yields

$$(C + D_\tau)^{-1} \sim \begin{pmatrix} C_I^{-1} & -C_I^{-1} C_{I,I^c} D_{\tau,I^c}^{-1} \\ -D_{\tau,I^c}^{-1} C_{I^c,I} C_I^{-1} & D_{\tau,I^c}^{-1} \end{pmatrix}$$

Plugging this in eq. (41), and again using that D_{τ,I^c}^{-1} is vanishingly small compared to constant matrices yields eq. (42). □

From the fact that by Lemma 3, $\tau D_{\tau,I} \sim$ const, it follows immediately that, if $I_t^c = \emptyset$,

$$(E_\tau)_{ij} = \begin{cases} \mathcal{O}(\tau) & i,j \in I^c \\ \text{const} & \text{otherwise} \end{cases} \tag{47}$$

For transition coordinates, eq. (40) may diverge or not, depending on the rate of $\hat{x}_{\tau,j} \to 0$. Define

$$J = I \cup \big\{j \in I_t^c \colon \lim_{\tau\to\infty} \tau^{\frac{1}{2}} \hat{x}_{\tau,j} \neq 0\big\}. \tag{48}$$

Then D_τ diverges on J^c and converges (but not necessarily vanishes) on J, and eqs. (45) and (46) remain valid if we use the set J rather than I to partition the matrix (with a small modification in eq. (45) to keep an extra possible constant term). Hence, we obtain the following modification of eq. (47):

$$(E_\tau)_{ij} = \begin{cases} \mathcal{O}(\tau) & i,j \in J^c \\ \text{const} & \text{otherwise} \end{cases} \tag{49}$$

B.5. **Asymptotic properties of the differential operator argument.** Next we consider the function $R_\tau(z)$ appearing in the argument of the differential operator in eq. (14) and defined in eq. (17),

$$\begin{aligned} R_\tau(z) &= \sum_{j=1}^{p} R_{\tau,j}(z_j) \\ R_{\tau,j}(z_j) &= \sum_{m\geq 3} \frac{1}{m}\Big[\frac{1}{(\mu-\hat{u}_{\tau,j})^m} + \frac{(-1)^m}{(\mu+\hat{u}_{\tau,j})^m}\Big](z_j)^m. \end{aligned}$$

We have the following result:

Lemma 4. *$R_{\tau,j}(z_j)$ is of the form*

$$R_{\tau,j}(z_j) = z_j^3 q_{\tau,j}(z_j)$$

with $q_{\tau,j}$ an analytic function in a region around $z_j = 0$ and

$$q_{\tau,j}(z_j) \leq \begin{cases} \mathcal{O}(\tau^2) & j \in J \\ \mathcal{O}(\tau) & j \in J^c \cap I_t^c \\ \mathit{const} & j \in I_{nt}^c \end{cases}$$

with J defined in eq. (48).

Proof. The first statement follows from the fact that the series expansion of $R_{\tau,j}(z_j)$ contains only powers of z_j greater than 3. The asymptotics as a function of τ for $j \in I$ and $j \in I_{nt}^c$ follow immediately from Lemma 3 and the definition of $R_{\tau,j}$ (Appendix B.2),

$$R_{\tau,j}(z_j) = -\ln\big[\mu^2 - (\hat{u}_{\tau,j}+z_j)^2\big] + \ln(\mu^2 - \hat{u}_{\tau,j}^2) - \frac{2\hat{u}_{\tau,j}}{\mu^2-\hat{u}_{\tau,j}^2}z_j - \frac{\mu^2+\hat{u}_{\tau,j}^2}{(\mu^2-\hat{u}_{\tau,j}^2)^2}z_j^2.$$

For $j \in J \cap I_t^c$, we have from Lemma 3 at worst $(\mu^2 - \hat{u}_{\tau,j}^2)^{-2} = \mathcal{O}\big[(\tau\hat{x}_{\tau,j})^2\big] \leq \mathcal{O}(\tau^2)$, whereas for $j \in J^c \cap I_t^c$, we have at worst $(\tau\hat{x}_{\tau,j})^2 = \tau(\tau^{\frac{1}{2}}\hat{x}_{\tau,j})^2 \leq \mathcal{O}(\tau)$. □

B.6. **Asymptotic approximation for the partition function.** To prove the analytic approximation eq. (18), we will show that successive terms in the series expansion of $e^{\frac{1}{4\tau^2}\Delta_\tau}$ result in terms of decreasing power in τ. The argument presented below is identical to existing proofs of the stationary phase approximation for multi-dimensional integrals (Wong, 2001), except that we need to track and estimate the dependence on τ in both Δ_τ and R_τ.

The series expansion of the differential operator exponential can be written as:

$$\begin{aligned} \exp\Big\{\frac{1}{4\tau^2}\Delta_\tau\Big\} &= \sum_{m\geq 0}\frac{1}{m!(2\tau)^{2m}}\Delta_\tau^m \\ &= \sum_{m\geq 0}\frac{1}{m!(2\tau)^{2m}}\sum_{j_1,\dots,j_{2m}=1}^{p} E_{j_1j_2}\dots E_{j_{2m-1}j_{2m}}\frac{\partial^{2m}}{\partial k_{j_1}\dots\partial k_{j_{2m}}} \\ &= \sum_{m\geq 0}\frac{1}{m!(2\tau)^{2m}}\sum_{\alpha\colon |\alpha|=2m} S_{\tau,\alpha}\frac{\partial^{2m}}{\partial k_1^{\alpha_1}\dots\partial k_p^{\alpha_p}}, \end{aligned}$$

where E is the matrix defined in eq. (41) (its dependence on τ is omitted for notational simplicity), $\alpha = (\alpha_1,\dots,\alpha_p)$ is a multi-index, $|\alpha| = \sum_j \alpha_j$, and $S_{\tau,\alpha}$ is the sum of all terms $E_{j_1j_2}\dots E_{j_{2m-1}j_{2m}}$ that give rise to the same multi-index α. From eq. (49), it follows that only coordinates in J^c give rise to diverging terms in $S_{\tau,\alpha}$, and only if they are coupled to other coordinates in J^c. Hence the total number $\sum_{j\in J^c}\alpha_j$ of J^c coordinates can be divided over at most $\frac{1}{2}\sum_{j\in J^c}\alpha_j$ E-factors, and we have

$$S_{\tau,\alpha} \leq \mathcal{O}\big(\tau^{\frac{1}{2}\sum_{j\in J^c}\alpha_j}\big).$$

Turning our attention to the partial derivatives, we may assume without loss of generality that the argument function Q is a finite sum of products of monomials and hence it is sufficient to prove eq. (18) with Q of the form $Q(z) = \prod_{j=1}^{p} Q_j(z_j)$. By Cauchy's theorem and Lemma 4, we have for $\epsilon > 0$ small enough,

$$\begin{aligned}
\frac{\partial^{\alpha_j}}{\partial k_j^{\alpha_j}} e^{R_{\tau,j}(ik_j)} Q_j(ik_j)\Big|_{k_j=0} &= \frac{\alpha_j!}{2\pi i}\oint_{|z|=\epsilon} \frac{1}{z^{\alpha_j+1}} e^{R_{\tau,j}(z_j)} Q_j(z_j)dz_j \\
&= \frac{\alpha_j!}{2\pi i}\sum_{n\geq 0}\frac{1}{n!}\oint_{|z|=\epsilon} z_j^{3n-\alpha_j-1} q_j(z_j)^n Q_j(z_j)dz_j \\
&= \frac{\alpha_j!}{2\pi i}\sum_{0\leq n<\frac{1}{3}(\alpha_j+1)}\frac{1}{n!}\oint_{|z|=\epsilon} z_j^{3n-\alpha_j-1} q_j(z_j)^n Q_j(z_j)dz \\
&\leq \begin{cases} \mathcal{O}\big(\tau^{\frac{2}{3}\alpha_j}\big) & j\in J \\ \mathcal{O}\big(\tau^{\frac{1}{3}\alpha_j}\big) & j\in J^c\cap I_t^c \\ \text{const} & j\in I_{nt}^c \end{cases}
\end{aligned}$$

The last result follows, because for $j\in J$ or $j\in J^c\cap I_t^c$, q_j scales at worst as τ^2 or τ, respectively, and hence, since only powers of q_j strictly less than $\frac{1}{3}(\alpha_j+1)$ contribute to the sum, the sum must be a polynomial in τ of degree less than $\frac{2}{3}\alpha_j$ or $\frac{1}{3}\alpha_j$, respectively (α_j can be written as either $3t$, $3t+1$ or $3t+2$ for some integer t; in all three cases, the largest integer strictly below $\frac{1}{3}(\alpha_j+1)$ equals t, and $t\leq\frac{1}{3}\alpha_j$).

Hence

$$\begin{aligned}
\sum_{\alpha\colon|\alpha|=2m} S_{\tau,\alpha}\frac{\partial^{2m}}{\partial k_1^{\alpha_1}\dots\partial k_p^{\alpha_p}} e^{R_\tau(ik)}Q(ik)\Big|_{k=0} &= \sum_{\alpha\colon|\alpha|=2m} S_{\tau,\alpha}\prod_j\frac{\partial^{\alpha_j}}{\partial k_j^{\alpha_j}} e^{R_{\tau,j}(ik_j)}Q_j(ik_j)\Big|_{k_j=0} \\
\leq \mathcal{O}\big(\tau^{\frac{1}{2}\sum_{j\in J^c}\alpha_j}\tau^{\frac{2}{3}\sum_{j\in J}\alpha_j+\frac{1}{3}\sum_{j\in J^c\cap I_t^c}\alpha_j}\big) &= \mathcal{O}\big(\tau^{\frac{2}{3}\sum_{j\in J}\alpha_j+\frac{1}{2}\sum_{j\in I_{nt}^c}\alpha_j+\frac{5}{6}\sum_{j\in J^c\cap I_t^c}\alpha_j}\big) \\
\leq \mathcal{O}\big(\tau^{\frac{5}{6}\sum_{j=1}^{p}\alpha_j}\big) = \mathcal{O}\big(\tau^{\frac{5}{3}m}\big)
\end{aligned}$$

This in turn implies that the m^{th} term in the expansion,

$$\exp\Big\{\frac{1}{4\tau^2}\Delta_\tau\Big\}e^{R_\tau(ik)}Q(ik)\Big|_{k=0} = \sum_{m\geq 0}\frac{1}{m!(2\tau)^{2m}}\sum_{\alpha\colon|\alpha|=2m} S_{\tau,\alpha}\prod_j\frac{\partial^{\alpha_j}}{\partial k_j^{\alpha_j}}e^{R_{\tau,j}(ik_j)}Q_j(ik_j)\Big|_{k_j=0} \qquad (50)$$

is bounded by a factor of $\tau^{-\frac{1}{3}m}$. Hence eq. (50) is an asymptotic expansion, with leading term

$$\exp\Big\{\frac{1}{4\tau^2}\Delta_\tau\Big\}e^{R_\tau(ik)}Q(ik)\Big|_{k=0} \sim \prod_{j=1}^{p}Q_j(0) = Q(0).$$

□

Appendix C. Zero-temperature limit of the partition function

The connection between the analytic approximation (18) and the minimum-energy (or maximum-likelihood) solution is established by first recalling that Fenchel's convex duality theorem implies that (Michoel, 2016)

$$\hat{x} = \operatorname*{argmin}_{x\in\mathbb{R}^p} H(x) = \operatorname*{argmin}_{x\in\mathbb{R}^p}\big[f(x)+g(x)\big] = \nabla f^*(-\hat{u}) = C^{-1}(w-\hat{u}),$$

where f and g are defined in eqs. (8)–(9),

$$f^*(u) = \max_{x\in\mathbb{R}^p}\bigl[x^T u - f(x)\bigr] = \frac{1}{2}(w+u)^T C^{-1}(w+u)$$

is the Legendre-Fenchel transform of f, and

$$\hat{u} = \operatorname*{argmin}_{\{u\in\mathbb{R}^p\colon |u_j|\leq\mu,\forall j\}} f^*(-u) = \operatorname*{argmin}_{\{u\in\mathbb{R}^p\colon |u_j|\leq\mu,\forall j\}} (w-u)^T C^{-1}(w-u). \tag{51}$$

One way of solving an optimization problem with constraints of the form $|u_j| \leq \mu$ is to approximate the hard constraints by a smooth, so-called 'logarithmic barrier function' (Boyd and Vandenberghe, 2004), i.e. solve the unconstrained problem

$$\hat{u}_\tau = \operatorname*{argmin}_{u\in\mathbb{R}^p}\Bigl[(w-u)^T C^{-1}(w-u) - \frac{1}{\tau}\sum_{j=1}^{p}\ln(\mu^2 - u_j^2)\Bigr] \tag{52}$$

such that in the limit $\tau\to\infty$, $\hat{u}_\tau \to \hat{u}$. Comparison with eqs. (12)–(13), shows that (52) is precisely the saddle point of the partition function, whereas the constrained optimization in eq. (51) was already encountered in eq. (37). Hence, let I again denote the set of non-zero coordinates in the maximum-likelihood solution $\hat{x}$. The following result characterizes completely the partition function in the limit $\tau\to\infty$, provided there are no transition coordinates.

Proposition 2. *Assume that μ is not a transition value, i.e. $j\in I \Leftrightarrow \hat{x}_j \neq 0 \Leftrightarrow |\hat{u}_j| = \mu$. Let $\sigma = \operatorname{sgn}(\hat{u})$ be the vector of signs of $\hat{u}$. Then $\operatorname{sgn}(\hat{x}_I) = \sigma_I$, and*

$$Z \sim \frac{e^{\tau(w_I-\mu\sigma_I)^T C_I^{-1}(w_I-\mu\sigma_I)}}{2^{\frac{|I|}{2}}\tau^{\frac{|I|}{2}+|I^c|}\sqrt{\det(C_I)}}\prod_{j\in I^c}\frac{\mu}{\mu^2-\hat{u}_j^2}. \tag{53}$$

In particular,

$$\lim_{\tau\to\infty}\frac{1}{\tau}\ln Z = (w_I-\mu\sigma_I)^T C_I^{-1}(w_I-\mu\sigma_I) = H(\hat{x}) = \min_{x\in\mathbb{R}^p} H(x).$$

Proof. First note that from the saddle point equations

$$(\mu^2 - \hat{u}_{\tau,j}^2)\hat{x}_{\tau,j} = \frac{\hat{u}_{\tau,j}}{\tau},$$

where as before $\hat{x}_\tau = C^{-1}(w-\hat{u}_\tau)$, and the fact that $|\hat{u}_{\tau,j}| < \mu$, it follows that $\operatorname{sgn}(\hat{x}_{\tau,j}) = \operatorname{sgn}(\hat{u}_{\tau,j})$ for all j and all τ. Let $j\in I$. Because $\hat{x}_{\tau,j}\to\hat{x}_j\neq 0$, it follows that there exists τ_0 large enough such that $\operatorname{sgn}(\hat{x}_{\tau,j}) = \operatorname{sgn}(\hat{x}_j)$ for all $\tau>\tau_0$. Hence also $\operatorname{sgn}(\hat{u}_{\tau,j}) = \operatorname{sgn}(\hat{x}_j)$ for all $\tau>\tau_0$, and since $\hat{u}_{\tau,j}\to\hat{u}_j\neq 0$, we must have $\operatorname{sgn}(\hat{u}_j) = \operatorname{sgn}(\hat{x}_j)$.

To prove eq. (53), we will calculate the leading term of $\det(C+D_\tau)$ in eq. (18). For this purpose, recall that for a square matrix M and any index subset I, we have (Horn and Johnson, 1985)

$$\det(M) = \det(M_I)\det(M_{I^c} - M_{I^cI}M_I^{-1}M_{II^c}) = \frac{\det(M_I)}{\det\bigl[(M^{-1})_{I^c}\bigr]} \tag{54}$$

Taking $M = C + D_\tau$, it follows from eqs. (43)–(46) that $\det(C_I + D_{\tau,I}) \sim \det(C_I)$, and $\det\bigl[(M^{-1})_{I^c}\bigr] \sim \det(D_{\tau,I^c}^{-1})$, and hence

$$\det(C+D_\tau) \sim \det(C_I)\det(D_{\tau,I^c}) = \tau^{|I^c|}\det(C_I)\prod_{j\in I^c}\frac{(\mu^2-\hat{u}_{\tau,j}^2)^2}{\mu^2+\hat{u}_{\tau,j}^2}.$$

Hence

$$\begin{aligned}\tau^{\frac{p}{2}}\prod_{j=1}^{p}\sqrt{\mu^2+\hat{u}_{\tau,j}^2}\sqrt{\det(C+D_\tau)} &\sim \tau^{\frac{p+|I^c|}{2}}\sqrt{\det(C_I)}\prod_{j\in I}\sqrt{\mu^2+\hat{u}_{\tau,j}^2}\prod_{j\in I^c}(\mu^2-\hat{u}_{\tau,j}^2)\\ &\sim \tau^{\frac{p+|I^c|}{2}}2^{\frac{|I|}{2}}\mu^{|I|}\sqrt{\det(C_I)}\prod_{j\in I^c}(\mu^2-\hat{u}_j^2),\end{aligned}$$

where the last line follows by replacing $\hat{u}_{\tau,j}$ by its leading term $\hat{u}_j$, and using $\hat{u}_j^2 = \mu^2$ for $j \in I$. Plugging this in eq. (18) and using eqs. (38)–(39) to get the leading term of the exponential factor results in eq. (53). □

The leading term in eq. (53) has a pleasing interpretation as a 'two-phase' system,

$$Z = \frac{1}{(2\pi)^{\frac{|I|}{2}}} Z_I Z_{I^c}$$

where Z_I and Z_{I^c} are the partition functions (normalization constants) of a multivariate Gaussian distribution and a product of independent shifted Laplace distributions, respectively:

$$\begin{aligned} Z_I &= \big(\frac{\pi}{\tau}\big)^{\frac{|I|}{2}} \frac{e^{\tau(w_I-\mu\sigma_I)^T C_I^{-1}(w_I-\mu\sigma_I)}}{\sqrt{\det(C_I)}} = \int_{\mathbb{R}^{|I|}} e^{-\tau[x_I^T C_I x_I - 2(w_I-\mu\sigma_I)^T x_I]} dx_I \\ Z_{I^c} &= \frac{1}{\tau^{|I^c|}} \prod_{j\in I^c} \frac{\mu}{\mu^2-\hat{u}_j^2} = \int_{\mathbb{R}^{|I^c|}} e^{-2\tau[\mu\sum_{j\in I^c}|x_j| - \hat{u}_{I^c}^T x_{I^c}]} dx_{I^c}. \end{aligned}$$

This suggests that in the limit $\tau \to \infty$, the non-zero maximum-likelihood coordinates are approximately normally distributed and decoupled from the zero coordinates, which each follow a shifted Laplace distribution. At finite values of τ however, this approximation is too crude, and more accurate results are obtained using the leading term of eq. (18). This is immediately clear from the fact that the partition function is a continous function of $w \in \mathbb{R}^p$, which remains true for the leading term of eq. (18), but not for eq. (53), which exhibits discontinuities whenever a coordinate enters or leaves the set I as w is smoothly varied.

Appendix D. Temperature-dependent optimization problem in coordinate space

The saddle point equations imply that $\hat{u}_\tau$ satisfies the convex optimization problem

$$\hat{u}_\tau = \operatorname*{argmin}_{x\in\mathbb{R}^p} \frac{1}{2}(w-u)^T C^{-1}(w-u) - \frac{1}{2\tau}\sum_{j=1}^{p} \ln(\mu^2 - u_j^2) = \operatorname*{argmin}_{x\in\mathbb{R}^p} f^*(-u) + g_\tau^*(u).$$

where $f^*(u) = \frac{1}{2}(w+u)^T C^{-1}(w+u)$ is the Legendre-Fenchel transform of $f(x) = \frac{1}{2}x^T Cx - w^T x$. The Legendre-Fenchel transform of g_τ^* is $g_\tau(x) = \sum_j g_{\tau,j}(x_j)$ with

$$g_{\tau,j}(x_j) = \max_{u_j\in\mathbb{R}} \Big(u_j x_j + \frac{1}{2\tau}\ln(\mu^2 - u_j^2) \Big).$$

Setting the derivative w.r.t. u_j to zero results in

$$x_j = \frac{u_j}{\tau(\mu^2 - u_j^2)}, \tag{55}$$

or

$$\tau x_j u_j^2 + u_j - \tau\mu^2 x_j = 0$$

with the solution that results in $|u_j| < \mu$ being

$$u_j = \frac{-1 + \sqrt{1 + 4\tau^2\mu^2 x_j^2}}{2\tau x_j}.$$

Hence

$$g_{\tau,j}(x_j) = \frac{\sqrt{1+4\tau^2\mu^2 x_j^2} - 1}{2\tau} + \frac{1}{2\tau}\ln\Big(\frac{\sqrt{1+4\tau^2\mu^2 x_j^2}}{2\tau^2 x_j^2}\Big).$$

By the saddle point equations, $\hat{x}_\tau = C^{-1}(w - \hat{u}_\tau)$ satisfies eq. (55) for $u = \hat{u}_\tau$, and hence $g_{\tau,j}(\hat{x}_{\tau,j}) = \hat{u}_{\tau,j}\hat{x}_{\tau,j} + \frac{1}{2\tau}\ln(\mu^2 - \hat{u}_{\tau,j}^2)$, or

$$\begin{aligned}
f(\hat{x}_\tau) + g_\tau(\hat{x}_\tau) &= \frac{1}{2}(w - \hat{u}_\tau)^T\hat{x}_\tau - w^T\hat{x}_\tau + \hat{u}_\tau^T\hat{x}_\tau + \frac{1}{2\tau}\sum_{j=1}^p \ln(\mu^2 - \hat{u}_{\tau,j}^2) \\
&= -\frac{1}{2}(w - \hat{u}_\tau)^T\hat{x}_\tau + \frac{1}{2\tau}\sum_{j=1}^p \ln(\mu^2 - \hat{u}_{\tau,j}^2) \\
&= -\big[f^*(-\hat{u}_\tau) + g^*(\hat{u}_\tau)\big] \\
&= -\min_{u\in\mathbb{R}^p}\big[f^*(-u) + g^*(u)\big] \\
&= \min_{x\in\mathbb{R}^p}\big[f(x) + g(x)\big],
\end{aligned}$$

where the last step uses Fenchel's convex duality theorem. This concludes the proof of eq. (21).

Appendix E. Analytic results for independent predictors

When predictors are independent, the matrix C is diagonal, and the partition function can be written as a product of one-dimensional integrals

$$Z = \int_{\mathbb{R}} e^{-\tau(cx^2 - 2wx + 2\mu|x|)}dx,$$

where $c, \mu > 0$ and $w \in \mathbb{R}$. This integral can be solved by writing $Z = Z^+ + Z^-$, where

$$\begin{aligned}
Z^\pm = \int_0^\infty e^{-\tau[cx^2 \pm 2(w\pm\mu)x]}dx = e^{\tau\frac{(w\pm\mu)^2}{c}}\int_0^\infty e^{-\tau c(x \pm \frac{w\pm\mu}{c})^2}dx = \frac{e^{\tau\frac{(w\pm\mu)^2}{c}}}{\sqrt{\tau c}}\int_{\pm\sqrt{\frac{\tau}{c}}(w\pm\mu)}^\infty e^{-y^2}dy \\
= \frac{1}{2}\sqrt{\frac{\pi}{\tau c}}e^{\tau\frac{(w\pm\mu)^2}{c}}\operatorname{erfc}\Big(\pm\sqrt{\frac{\tau}{c}}(w\pm\mu)\Big) = \frac{1}{2}\sqrt{\frac{\pi}{\tau c}}\operatorname{erfcx}\Big(\pm\sqrt{\frac{\tau}{c}}(w\pm\mu)\Big), \quad (56)
\end{aligned}$$

where $\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-y^2}dy$ and $\operatorname{erfcx}(x) = e^{x^2}\operatorname{erfc}(x)$ are the complementary and scaled complementary error functions, respectively. Hence,

$$\log Z = \log\Big[\operatorname{erfcx}\Big(\sqrt{\frac{\tau}{c}}(\mu + w)\Big) + \operatorname{erfcx}\Big(\sqrt{\frac{\tau}{c}}(\mu - w)\Big)\Big] + \frac{1}{2}\big(\log\pi - \log(\tau c)\big) - \log 2,$$

and

$$\begin{aligned}
\hat{x}_\tau = \mathbb{E}(x) &= \frac{1}{2\tau}\frac{\partial \log Z}{\partial w} \\
&= \frac{1}{c}\frac{(\mu + w)\operatorname{erfcx}\big(\sqrt{\frac{\tau}{c}}(\mu + w)\big) - (\mu - w)\operatorname{erfcx}\big(\sqrt{\frac{\tau}{c}}(\mu - w)\big)}{\operatorname{erfcx}\big(\sqrt{\frac{\tau}{c}}(\mu + w)\big) + \operatorname{erfcx}\big(\sqrt{\frac{\tau}{c}}(\mu - w)\big)} \\
&= \frac{w}{c} + \frac{\mu}{c}\frac{\operatorname{erfcx}\big(\sqrt{\frac{\tau}{c}}(\mu + w)\big) - \operatorname{erfcx}\big(\sqrt{\frac{\tau}{c}}(\mu - w)\big)}{\operatorname{erfcx}\big(\sqrt{\frac{\tau}{c}}(\mu + w)\big) + \operatorname{erfcx}\big(\sqrt{\frac{\tau}{c}}(\mu - w)\big)} \\
&= \frac{w}{c} + (1 - 2\alpha)\frac{\mu}{c},
\end{aligned}$$

where

$$\alpha = \frac{1}{1 + \frac{\operatorname{erfcx}\big(\sqrt{\frac{\tau}{c}}(\mu - w)\big)}{\operatorname{erfcx}\big(\sqrt{\frac{\tau}{c}}(\mu + w)\big)}}.$$

APPENDIX F. NUMERICAL RECIPES

F.1. **Solving the saddle point equations.** To calculate the partition function and posterior distribution at any value of τ, we need to solve the set of equations in eq. (13). To avoid having to calculate the inverse matrix C^{-1}, we make a change of variables $x = C^{-1}(w-u)$, or $u = w - Cx$, such that eq. (13) becomes

$$x_j\big[w_j - (Cx)_j - \mu\big]\big[w_j - (Cx)_j + \mu\big] + \frac{1}{\tau}\big[w_j - (Cx)_j\big] = 0. \tag{57}$$

We will use a coordinate descent algorithm where one coordinate of x is updated at a time, using the current estimates $\hat{x}$ for the other coordinates. Defining

$$a_j = w_j - \sum_{k\neq j} C_{kj}\hat{x}_k,$$

we can write eq. (57) as

$$C_{jj}^2 x_j^3 - 2a_j C_{jj} x_j^2 + \big(a_j^2 - \mu^2 - \frac{C_{jj}}{\tau}\big)x_j + \frac{a_j}{\tau} = 0$$

The roots of this 3rd order polynomial are easily obtained numerically, and by construction there will be a unique root for which $u_j = w_j - (Cx)_j = a_j - C_{jj}x_j$ is located in the interval $(-\mu,\mu)$. This root will be the new estimate $\hat{x}_j$. Given a new $\hat{x}_j^{(\text{new})}$, we can update the vector a as

$$a_k^{(\text{new})} = \begin{cases} a_j^{(\text{old})} & k = j \\ a_k^{(\text{old})} - C_{kj}\big(\hat{x}_j^{(\text{new})} - \hat{x}_j^{(\text{old})}\big) & k \neq j \end{cases}$$

and proceed to update the next coordinate.

After all coordinates of $\hat{x}$ have converged, we obtain $\hat{u}_\tau$ by performing the matrix-vector operation

$$\hat{u}_\tau = w - C\hat{x},$$

or, if we only need the expectation values,

$$\mathbb{E}_\tau(x) = \hat{x}.$$

For $\tau = \infty$, the solution to eq. (57) is given by the maximum-likelihood effect size vector (cf. Appendix C), for which ultra-fast algorithms exploiting the sparsity of the solution are available (Friedman et al., 2010). Hence we use this vector as the initial vector for the coordinate descent algorithm for $\tau < \infty$ and expect fast convergence if τ is large. Solutions for multiple values of τ can be obtained along a descending path of τ-values, each time taking the previous solution as the initial vector for finding the next solution.

F.2. **High-dimensional determinants in the partition function.** Calculating the stationary phase approximation to the partition function involves the computation of the p-dimensional determinant $\det(C + D_\tau)$ [cf. eq. (18)], which can become computationally expensive in high-dimensional settings. However, when C is of the form $C = \frac{A^T K^{-1} A}{2n} + \lambda\mathbb{1}$ [cf. eq. (6)] with $A \in \mathbb{R}^{n\times p}$, $K \in \mathbb{R}^{n\times n}$ invertible, and $p > n$, these determinants can be written as n-dimensional determinants, using the matrix determinant lemma:

$$\det(C + D_\tau) = \det\Big(\frac{A^T K^{-1} A}{2n} + D'_\tau\Big) = \frac{\det(D'_\tau)}{\det(K)}\det\Big(K + \frac{A(D'_\tau)^{-1}A^T}{2n}\Big), \tag{58}$$

where $D'_\tau = D_\tau + \lambda\mathbb{1}$ is a diagonal matrix whose determinant and inverse are trivial to obtain.

To avoid numerical overflow or underflow, all calculations are performed using logarithms of partition functions. For n large, a numerically stable computation of eq. (58) uses the equality $\log\det B = \operatorname{tr}\log B = \sum_{i=1}^n \log\epsilon_i$, where $B = K + \frac{1}{2n}A(D'_\tau)^{-1}A^T$ and ϵ_i are the eigenvalues of B.

F.3. **Marginal posterior distributions.** Calculating the marginal posterior distributions $p(x_j)$ [eq. 22] requires applying the analytic approximation eq. (14) using a different $\hat{u}_\tau$ for every different value of x_j. To make this process more efficient, two simple properties are exploited:

(1) For $x_j = \hat{x}_{\tau,j}$, the saddle point for the $(p-1)$-dimensional partition function $Z(C_{I_j}, w_{I_j} - x_j C_{j,I_j}, \mu)$ is given by the original saddle point vector $\hat{x}_{\tau,k}$, $k \neq j$. This follows easily from the saddle point equations.

(2) If x_j changes by a small amount, the new saddle point also changes by a small amount. Hence, taking the current saddle point vector for x_j as the starting vector for solving the set of saddle point equations for the next value $x_j + \delta$ results in rapid convergence (often in a single loop over all coordinates).

Hence we always start by computing $p(x_j = \hat{x}_{\tau,j})$ and then compute $p(x_j)$ separately for a series of ascending values $x_j > \hat{x}_{\tau,j}$ and a series of descending values $x_j < \hat{x}_{\tau,j}$

F.4. **Sampling from the one-dimensional distribution.** Consider again the case of one predictor, with posterior distribution

$$p(x) = \frac{e^{-\tau(cx^2 - 2wx + 2\mu|x|)}}{Z}. \tag{59}$$

To sample from this distribution, note that

$$p(x) = (1-\alpha)\, p(x \mid x < 0) + \alpha\, p(x \mid x \geq 0),$$

where

$$p(x \mid x \in \mathbb{R}^{\pm}) = \frac{e^{-\tau(cx^2 - 2(w \mp \mu)x)}}{Z^{\mp}}, \tag{60}$$

$Z^{\pm}$ were defined in eq. (56), and

$$\alpha = P(x \geq 0) = \int_0^\infty p(x) dx = \frac{1}{Z} \int_0^\infty e^{-\tau[cx^2 - 2(w-\mu)x]} dx = \frac{Z^-}{Z} = \frac{1}{1 + \frac{\operatorname{erfcx}\left(\sqrt{\frac{\tau}{c}}(\mu - w)\right)}{\operatorname{erfcx}\left(\sqrt{\frac{\tau}{c}}(\mu + w)\right)}}.$$

Eq. (60) defines two truncated normal distributions with means $(w \mp \mu)/c$ and standard deviation $1/\sqrt{2\tau c}$, for which sampling functions are available. Hence, to sample from the distribution (59), we first sample a Bernoulli random variable with probability α, and then sample from the appropriate truncated normal distribution.

F.5. **Gibbs sampler.** To sample from the Gibbs distribution in the general case, we use the 'basic Gibbs sampler' of Hans (2011). Let $\hat{x}$ be the current vector of sampled regression coefficients. Then a new coefficient x_j is sampled from the conditional distribution

$$p(x_j \mid \{\hat{x}_k, k \neq j\}) = \frac{e^{-\tau[C_{jj}x_j^2 - 2a_j x_j + 2\mu|x_j|]}}{Z_j}, \tag{61}$$

where $a_j = w_j - \sum_{k \neq j} C_{kj} \hat{x}_k$ and Z_j is a normalization constant. This distribution is of the same form as eq. (59) and hence can be sampled from in the same way. Notice that, as in section F.1, after sampling a new $\hat{x}_j$, we can update the vector a as

$$a_k^{(\text{new})} = \begin{cases} a_j^{(\text{old})} & k = j \\ a_k^{(\text{old})} - C_{kj}\left(\hat{x}_j^{(\text{new})} - \hat{x}_j^{(\text{old})}\right) & k \neq j \end{cases}.$$

F.6. **Maximum a-posteriori estimation of the inverse temperature.** This paper is concerned with the problem of obtaining the posterior regression coefficient distribution for the Bayesian lasso and elastic net when values for the hyperparameters (λ, μ, τ) are given. There is abundant literature on how to select values for λ and μ for maximum-likelihood estimation, mainly through cross validation or by predetermining a specific level of sparsity (i.e. number of non-zero predictors). Hence we assume an appropriate choice for λ and μ has been made, and propose to then set τ equal to a first-order approximation of its maximum a posteriori (MAP) value, i.e. finding the value which maximizes the log-likelihood of observing data $y \in \mathbb{R}^n$ and $A \in \mathbb{R}^p$, similar to what was suggested by Hans (2011). To do so we must include the normalization constants in the prior distributions (2)–(3):

$$\begin{aligned} p(y \mid A, x, \tau) &= \Bigl(\frac{\tau}{2\pi n}\Bigr)^{\frac{n}{2}} e^{-\frac{\tau}{2n}\|y-Ax\|^2} = \Bigl(\frac{\tau}{2\pi n}\Bigr)^{\frac{n}{2}} e^{-\frac{\tau}{2n}\|y\|^2} e^{-\frac{\tau}{2n}[x^T A^T Ax - 2(A^T y)^T x]} \\ p(x \mid \lambda, \mu, \tau) &= \frac{e^{-\tau(\lambda\|x\|^2 + 2\mu\sum_j |x_j|)}}{Z_0} \end{aligned}$$

where for $\lambda > 0$,

$$\begin{aligned} Z_0 &= \int_{\mathbb{R}^p} dx\, e^{-\tau(\lambda\|x\|^2+2\mu\sum_j |x_j|)} = \Bigl(\int_{\mathbb{R}} dx\, e^{-\tau(\lambda x^2 + 2\mu|x|)}\Bigr)^p = \Bigl(2\int_0^\infty dx\, e^{-\tau(\lambda x^2+2\mu x)}\Bigr)^p \\ &= \Bigl(\frac{2e^{\frac{\mu^2\tau}{\lambda}}}{\sqrt{\lambda\tau}} \int_{\sqrt{\frac{\mu^2\tau}{\lambda}}}^\infty e^{-t^2} dt\Bigr)^p = \Bigl(\sqrt{\frac{\pi}{\lambda\tau}} e^{\frac{\mu^2\tau}{\lambda}} \operatorname{erfc}\Bigl(\sqrt{\frac{\mu^2\tau}{\lambda}}\Bigr)\Bigr)^p \sim \Bigl(\frac{1}{\mu\tau}\Bigr)^p, \qquad (62) \end{aligned}$$

and the last relation follows from the first-order term in the asymptotic expansion of the complementary error function for large values of its argument,

$$\operatorname{erfc}(x) \sim \frac{e^{-x^2}}{x\sqrt{\pi}}.$$

For pure lasso regression ($\lambda = 0$), this relation is exact:

$$Z_0 = \Bigl(\frac{1}{\mu\tau}\Bigr)^p.$$

Hence, the log-likelihood of observing data $y \in \mathbb{R}^n$ and $A \in \mathbb{R}^p$ given values for λ, μ, τ is

$$\begin{aligned} \mathcal{L} &= \log \int_{\mathbb{R}^p} dx\, p(y \mid A, x, \tau) p(x \mid \lambda, \mu, \tau) \\ &= \frac{n}{2}\log\tau - \frac{\|y\|^2}{2n}\tau - \log Z_0 + \log \int_{\mathbb{R}^p} dx\, e^{-\tau H(x)} + \text{const}, \end{aligned}$$

where 'const' are constant terms not involving the hyperparameters. Taking the first order approximation

$$\log Z = \log \int_{\mathbb{R}^p} dx\, e^{-\tau H(x)} \sim -\tau H_{\min} = -\tau H(\hat{x}),$$

where $\hat{x}$ are the maximum-likelihood regression coefficients, we obtain

$$\begin{aligned} \mathcal{L} &\sim \bigl(p + \frac{n}{2}\bigr)\log\tau - \Bigl[\frac{\|y\|^2}{2n} + H(\hat{x})\Bigr]\tau + p\log\mu \\ &= \bigl(p + \frac{n}{2}\bigr)\log\tau - \Bigl[\frac{1}{2n}\|y - A\hat{x}\|^2 + \lambda\|\hat{x}\|^2 + 2\mu\|\hat{x}\|_1\Bigr]\tau + p\log\mu \end{aligned}$$

which is maximized at

$$\tau = \frac{p + \frac{n}{2}}{\frac{1}{2n}\|y - A\hat{x}\|^2 + \lambda\|\hat{x}\|^2 + 2\mu\|\hat{x}\|_1}.$$

Note that a similar approach to determine the MAP value for λ would require keeping an additional second order term in eq. (62), and that for $p > n$ it is not possible to simultaneously determine MAP values for all three hyperparameters, because it leads to a set of equations that are solved by the combination $\lambda = \mu = 0$ and $\tau = \infty$.

APPENDIX G. EXPERIMENTAL DETAILS

G.1. **Hardware and software.** All numerical experiments were performed on a standard Macbook Pro with 1.8 GHz processor and 16 GB RAM running macOS version 10.12.6 and Matlab version R2017a. Maximum-likelihood elastic net models were fitted using Glmnet for Matlab (`https://web.stanford.edu/~hastie/glmnet_matlab/`). Matlab software to solve the saddle point equations, compute the partition function and marginal posterior distributions, and run a Gibbs sampler, are available at `https://github.com/tmichoel/bayonet/`. This site also contains copies of the test data sets and scripts to reproduce the figures from this paper.

G.2. **Independent predictors.** For the analysis in Figure 2, parameter values were set to $p = 1$, $C = 1.0$, $w = 0.5$, μ ranging from 0.05 to 5 in 20 geometric steps, and τ ranging from 10 to 10^9 in 33 geometric steps.

G.3. **Diabetes and leukemia data.** The diabetes data were obtained from `https://web.stanford.edu/~hastie/CASI_files/DATA/diabetes.html`. The leukemia data were obtained from `https://web.stanford.edu/~hastie/CASI_files/DATA/leukemia.html`. Data were standardized according to eq. (1), and no further processing was performed. For the analysis in Figure 2, parameter values were set to $\lambda = 0.1$, μ ranging from $0.01\mu_{\max}$ upto, but not including, $\mu_{\max} = \max_j |w_j|$ in 20 geometric steps, and τ ranging from 10 to 10^9 in 33 geometric steps. For the analysis in Figure 3, λ was set to 0.1, μ was selected as the smallest value with a maximum-likelihood solution with 5 (diabetes data) or 10 (leukemia data) non-zero predictors, and τ was set to its maximum a-posteriori value given λ and μ [$\tau = 682.3$ (diabetes data) and $9.9439 \cdot 10^3$ (leukemia data].

G.4. **Cancer Cell Line Encyclopedia data.** Normalized expression data for 18,926 genes in 917 cancer cell lines were obtained from the Gene Expression Omnibus accession number GSE36139 using the Series Matrix File `GSE36139-GPL15308_series_matrix.txt`. Drug sensitivity data for 24 compounds in 504 cell lines were obtained from the supplementary material of Barretina et al. (2012) (tab 11 from supplementary file `nature11003-s3.xls`); 474 cell lines were common between gene expression and drug response data and used for our analyses. Of the available drug response data, only the activity area ('actarea') variable was used; 7 compounds had more than 40 zero activity area values (meaning inactive compounds) in the 474 cell lines and were discarded. For the remaining 17 compounds, the following procedure was performed:

(1) Hyper-parameters were set to $\lambda = 0.1$; $\mu_n = \mu_{\max} \times r^{\frac{N+1-n}{N}}$, where $N = 20$, $n = 1, 2, \dots, 20$, $r = 0.01$ and $\mu_{\max} = \max_{j=1,\dots,p} |w_j|$, with w as defined in eq. (6)–(7) and $p = 18,926$; $\tau_m = 10^{0.25(m+M-1)}$, where $M = 12$, $m = 1, 2, \dots, 13$.
(2) 470 randomly selected cell lines were randomly divided in 10 sets of 47 samples. Each set was used to validate predictions of models trained on the remaning 423 samples.
(3) For each training data set, and for each drug, the following procedure was performed:
  (a) The 1,000 genes most strongly correlated with the response were selected as candidate predictors.
  (b) Response and predictor data were standardized.
  (c) Maximum-likelihood coefficients for ridge regression ($\mu = 0$) and elastic net regression for each μ_n were calculated.
  (d) Bayesian posterior expectation values for each μ_n and each τ_m were calculated.
  (e) Drug responses were predicted on the original data scale in the 47 held-out validation samples using all sets of regression coefficients, and the Pearson correlation with the true drug response was calculated.
  (f) For each drug, each value of μ and each value of τ, the median correlation value over the 10 predictions was taken, resulting in a single value for ridge regression, 20 values for maximum-likelihood elastic net regression, and 13×20 values for Bayesian elastic net regression.

The top 1,000 most correlated genes were pre-filtered in each training data set, partly because in trial runs this resulted in better predictive performance than pre-selecting 5,000 or 10,000 genes, and partly to speed up calculations.

Appendix H. Supplementary figures

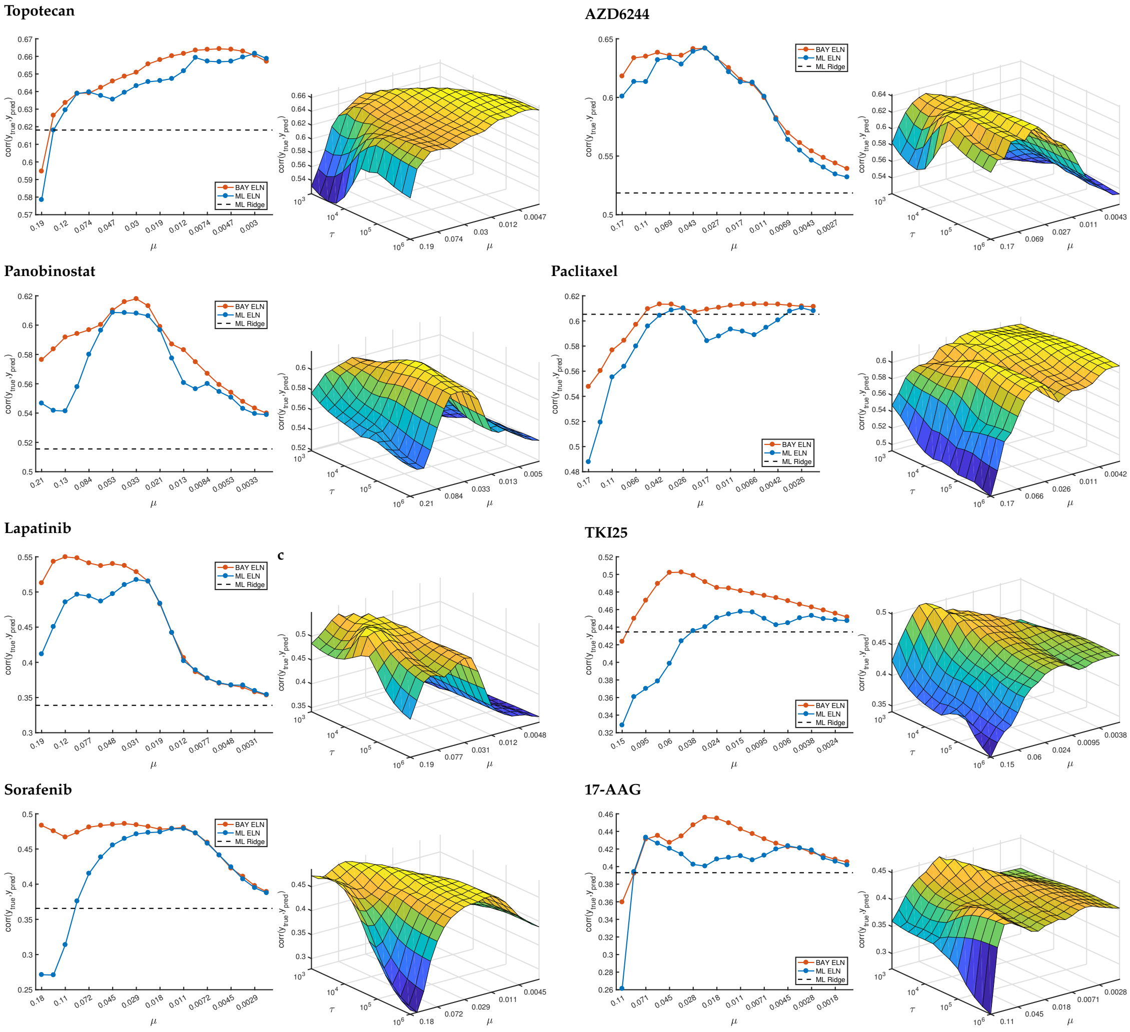

Figure S1. Same as Figure 4b and c, for drugs 2–9 from Figure 4a.

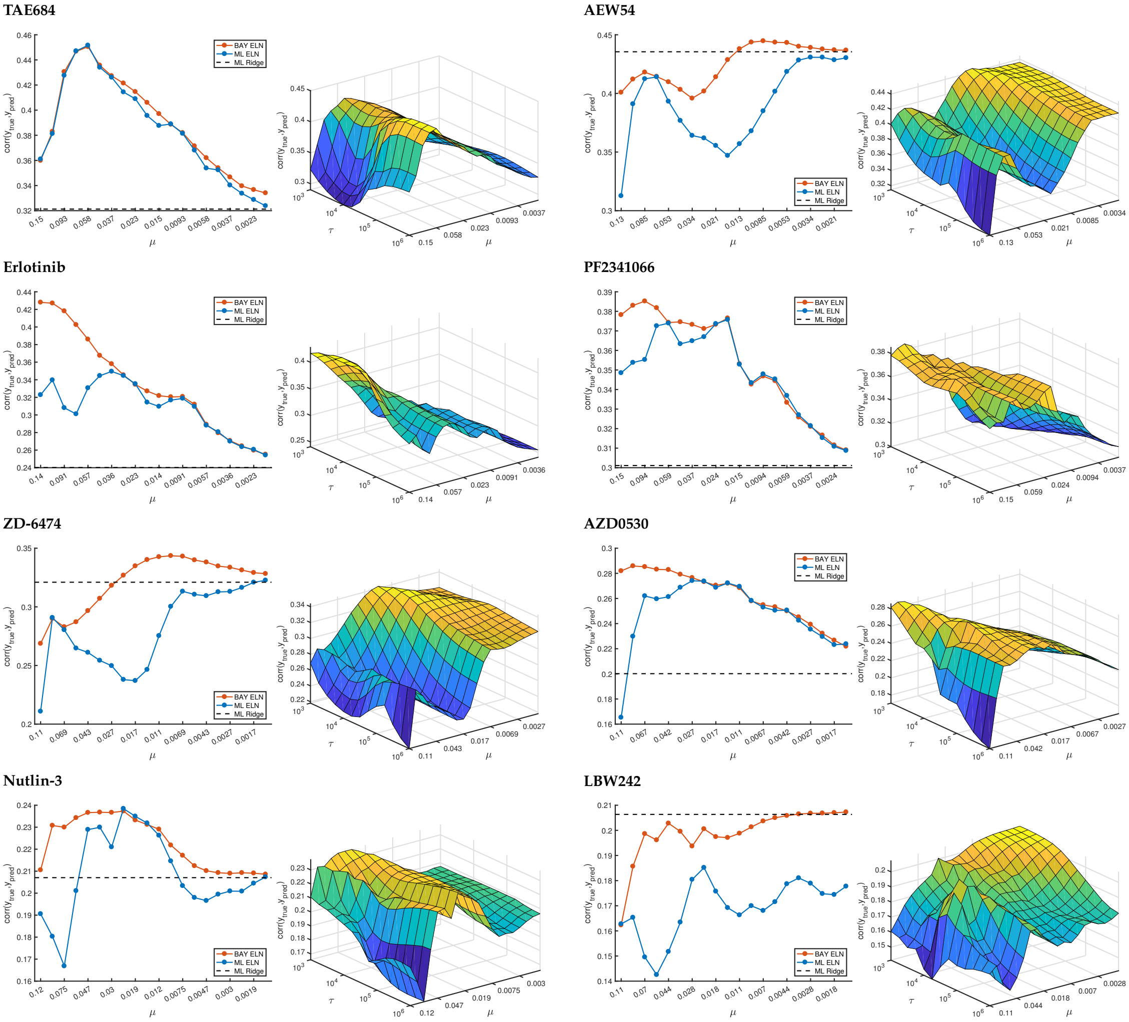

FIGURE S2. Same as Figure 4b and c, for drugs 10–17 from Figure 4a.

References

Alonso, M. and G. Forbes (1995). Fractional Legendre transformation. *Journal of Physics A: Mathematical and General 28*(19), 5509.

Barretina, J., G. Caponigro, N. Stransky, K. Venkatesan, A. Margolin, S. Kim, C. Wilson, J. Lehár, G. Kryukov, D. Sonkin, et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature 483*(7391), 603–607.

Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, 631–650.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–499.

El Ghaoui, L., V. Viallon, and T. Rabbani (2012). Safe feature elimination for the lasso and sparse supervised learning problems. *Pacific Journal of Optimization 8*, 667–698.

Fedoryuk, M. V. (1971). The stationary phase method and pseudodifferential operators. *Russian Mathematical Surveys 26*(1), 65–115.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*. Springer series in statistics Springer, Berlin.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 835–845.

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association 106*(496), 1383–1393.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Horn, R. A. and C. R. Johnson (1985). *Matrix analysis*. Cambridge University Press.

Hunter, J. K. and B. Nachtergaele (2001). *Applied analysis*. World Scientific Publishing Co Inc.

Kass, R. E. and D. Steffey (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association 84*(407), 717–726.

Lang, S. (2002). *Complex Analysis*, Volume 103 of *Graduate Texts in Mathematics*. Springrer.

Lee, J. S., C. Kwon, and H. Park (2013). Modified saddle-point integral near a singularity for the large deviation function. *Journal of Statistical Mechanics: Theory and Experiment 2013*(11), P11002.

Li, Q., N. Lin, et al. (2010). The Bayesian elastic net. *Bayesian Analysis 5*(1), 151–170.

Litvinov, G. L. (2005). The Maslov dequantization, idempotent and tropical mathematics: a brief introduction. *arXiv preprint math/0507014*.

Liu, J. S. (2004). *Monte Carlo strategies in scientific computing*. Springer.

Mallick, H. and N. Yi (2013). Bayesian methods for high dimensional linear models. *Journal of Biometrics & Biostatistics 1*, 005.

Michoel, T. (2016). Natural coordinate descent algorithm for L1-penalised regression in generalised linear models. *Computational Statistics & Data Analysis 97*, 60–70.

Osborne, M., B. Presnell, and B. Turlach (2000a). On the lasso and its dual. *Journal of Computational and Graphical Statistics 9*(2), 319–337.

Osborne, M. R., B. Presnell, and B. A. Turlach (2000b). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis 20*(3), 389–403.

Park, T. and G. Casella (2008). The Bayesian lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Rajaratnam, B. and D. Sparks (2015a). Fast Bayesian lasso for high-dimensional regression. *arXiv preprint arXiv:1509.03697*.

Rajaratnam, B. and D. Sparks (2015b). MCMC-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains. *arXiv preprint arXiv:1508.00947*.

Rakitsch, B., C. Lippert, O. Stegle, and K. Borgwardt (2012). A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics 29*(2), 206–214.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 319–392.

Schneidemann, V. (2005). *Introduction to Complex Analysis in Several Variables*. Birkhäuser Verlag.

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

Tibshirani, R. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics 7*, 1456–1490.

Tibshirani, R., J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. Tibshirani (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(2), 245–266.

Wong, R. (2001). *Asymptotic approximation of integrals*, Volume 34. SIAM.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.

THE ROSLIN INSTITUTE, THE UNIVERSITY OF EDINBURGH, EASTER BUSH, MIDLOTHIAN, EH25 9RG, UK
E-mail address: tom.michoel@roslin.ed.ac.uk