Regularization with variance-mean mixtures

Nick Polson
University of Chicago

James Scott
University of Texas at Austin

Workshop on Sensing and Analysis of High-Dimensional Data
Duke University
July 2011

Robust regression	Student t
Logit models	Lasso/ridge/bridge
Multinomial logit models	MC+
Extreme-value models	Group lasso
Support vector machines	Normal/exponential-gamma
Topic models	Normal/gamma
Restricted Boltzmann machines	Generalized double Pareto
Neural networks	Normal/inverted-beta
Autologistic models	Normal/inverse-Gaussian
Penalized additive models	Meridian filter

Robust regression	Student t
Logit models	Lasso/ridge/bridge
Multinomial logit models	MC+
Extreme-value models	Group lasso
Support vector machines	Normal/exponential-gamma
Topic models	Normal/gamma
Restricted Boltzmann machines	Generalized double Pareto
Neural networks	Normal/inverted-beta
Autologistic models	Normal/inverse-Gaussian
Penalized additive models	Meridian filter

Our approach works for arbitrary combinations of likelihood with prior.

No matrix inversion; no numerical derivatives.

Fully parallelizable block updates.

Robust regression	Student t
Logit models	Lasso/ridge/bridge
Multinomial logit models	MC+
Extreme-value models	Group lasso
Support vector machines	Normal/exponential-gamma
Topic models	Normal/gamma
Restricted Boltzmann machines	Generalized double Pareto
Neural networks	Normal/inverted-beta
Autologistic models	Normal/inverse-Gaussian
Penalized additive models	Meridian filter

Our approach works for arbitrary combinations of likelihood with prior.

No matrix inversion; no numerical derivatives.

Fully parallelizable block updates.

"But I can solve all of these with an ℓ^1 constraint."

True enough.

But consider an example: p = 25; n = 500; 1000 simulated data sets.

$$
\begin{aligned}
y_i &= 1_{z_i>0} \text{ for } i = 1, \ldots, n \\
\mathbf{z} &\sim \mathrm{N}(X\beta, I) \\
\beta &= \begin{cases} \sqrt{5} & (5x) \\ 0 & (20x) \end{cases}
\end{aligned}
$$

	MLE	Lasso-UT	Lasso-CV	HS
Median SSE	19.0	15.3	12.3	0.7
Mean SSE	68.6	15.4	11.7	1.6

“But I can solve all of these with an ℓ^1 constraint.”

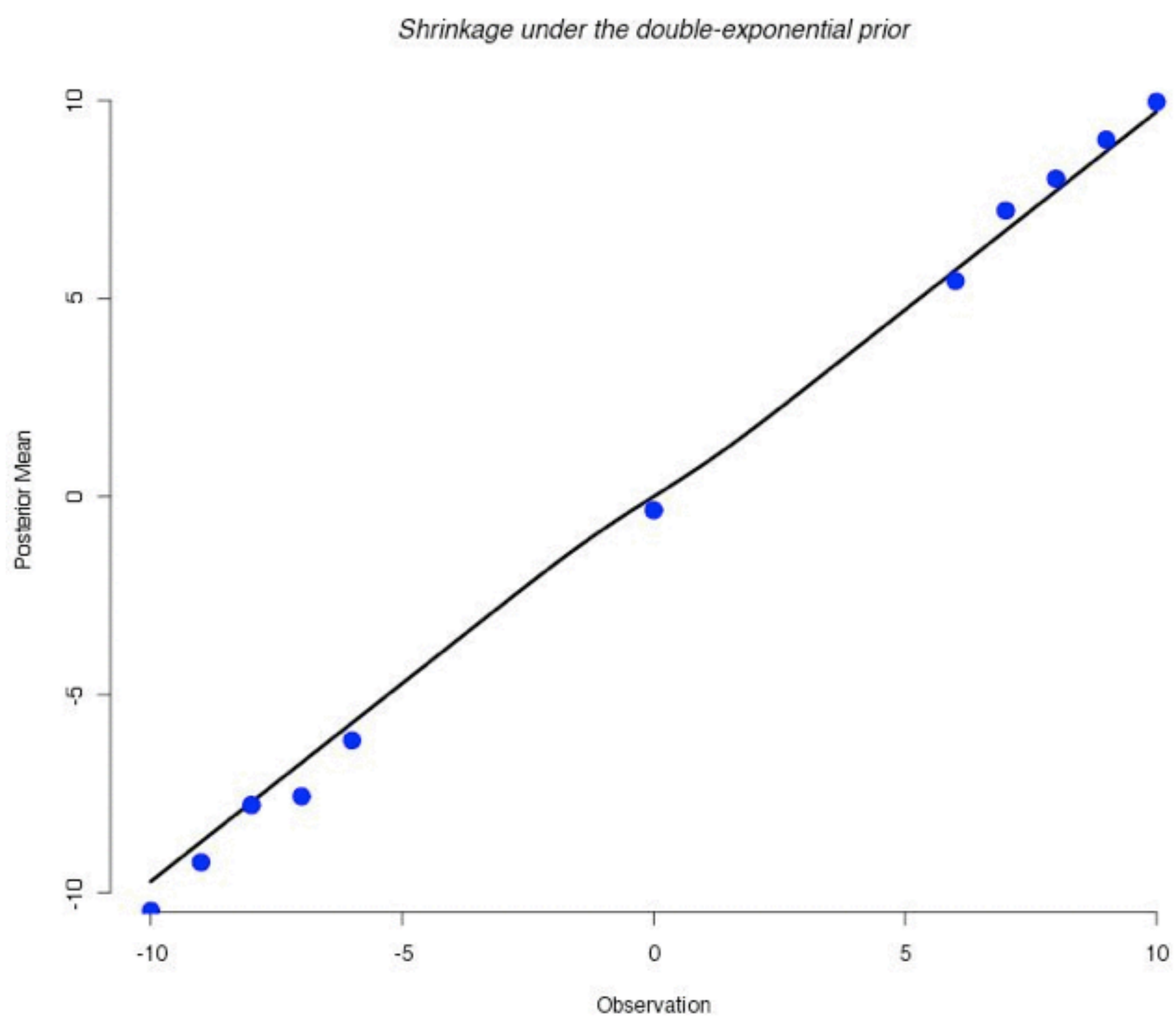
True enough.

But consider an example: p = 25; n = 500; 1000 simulated data sets.

$$
\begin{aligned}
y_i &= 1_{z_i > 0} \text{ for } i = 1, \ldots, n \\
\mathbf{z} &\sim \mathrm{N}(X\beta, I) \\
\beta &= \begin{cases} \sqrt{5} & (5x) \\ 0 & (20x) \end{cases}
\end{aligned}
$$

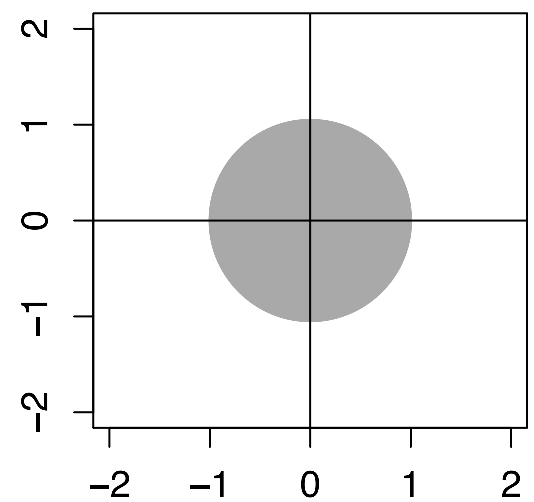
(an “r-spike” signal)

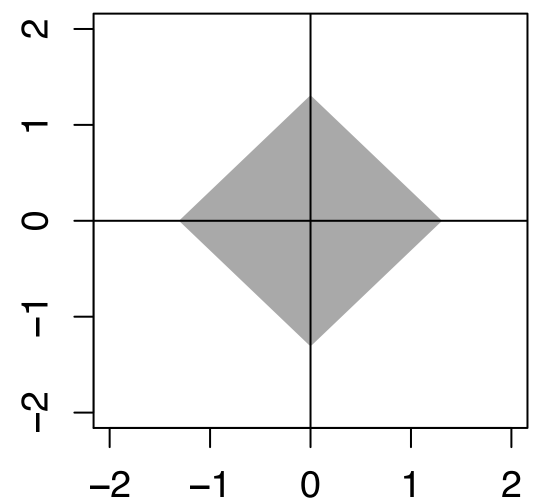
	MLE	Lasso-UT	Lasso-CV	HS
Median SSE	19.0	15.3	12.3	0.7
Mean SSE	68.6	15.4	11.7	1.6



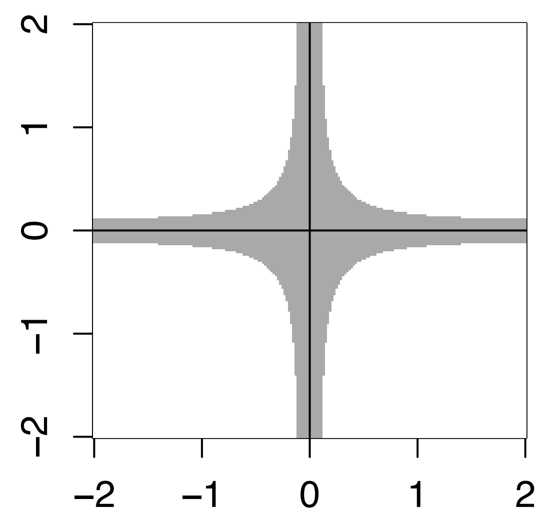
Shrinkage under the double-exponential prior
Posterior Mean
Observation
-10
-5
0
5
10

There seems to be precise separation of the computationally nice penalties/priors . . .



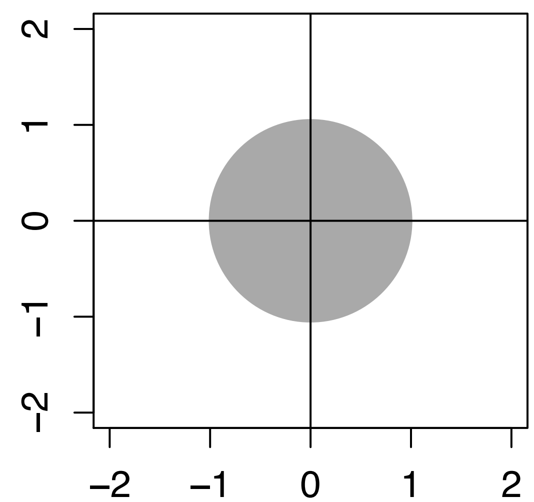


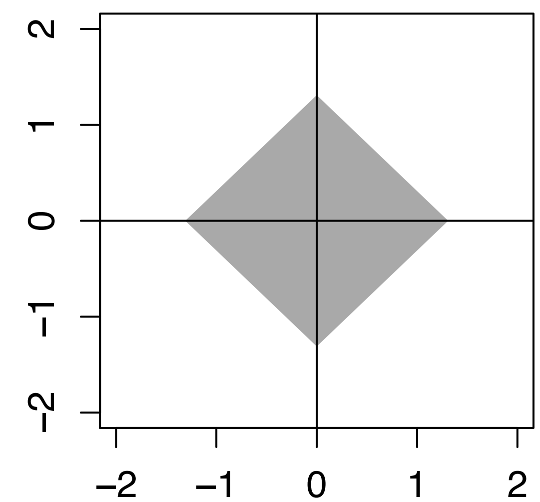
. . . from the "statistically nice" priors.



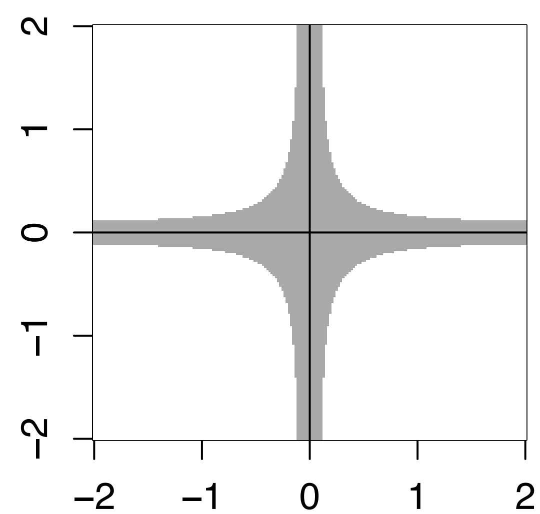
Our contribution: an algorithm for many priors in this second class, when used in conjunction with many common non-Gaussian likelihoods.

There seems to be precise separation of the computationally nice penalties/priors . . .





. . . from the "statistically nice" priors.



PS, 2011
Fan and Li, 2001
Pericchi and Smith, 1992
Masreliez, 1975
Brown, 1971
etc.

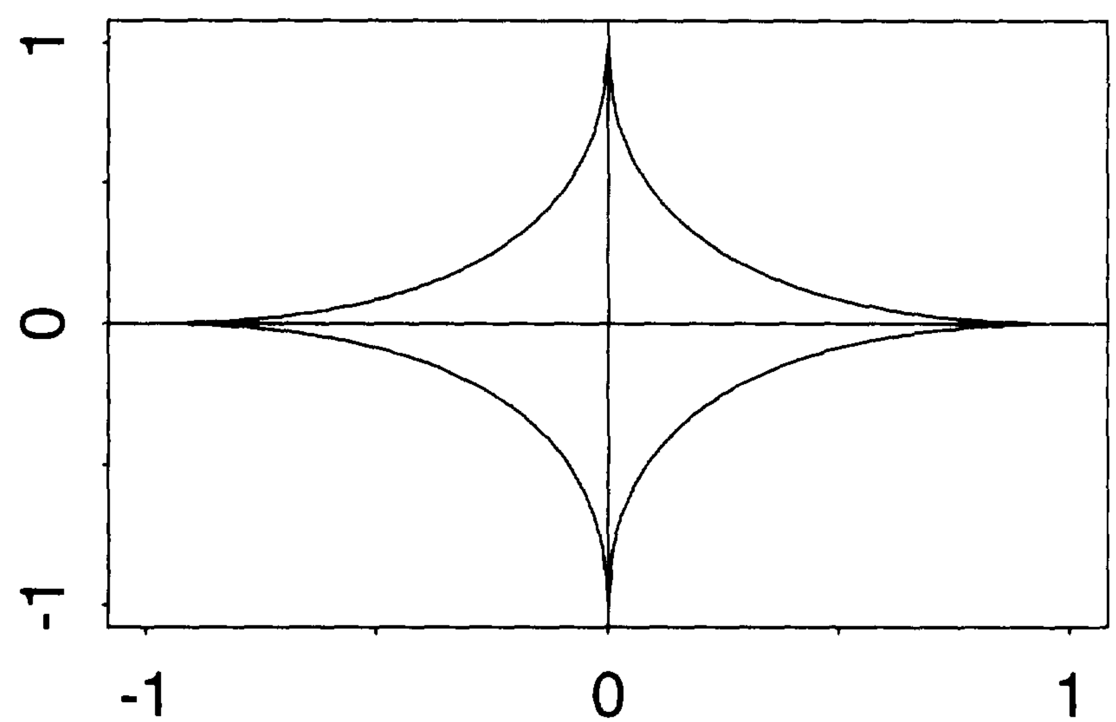
Our contribution: an algorithm for many priors in this second class, when used in conjunction with many common non-Gaussian likelihoods.

A teaser example: logit with a bridge penalty

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \log(1 + \exp\{-y_i x_i^T \beta\}) + \sum_{j=1}^{p} |\beta_j / \tau s_j|^{\alpha} \right)$$

A teaser example: logit with a bridge penalty

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \log(1 + \exp\{-y_i x_i^T \beta\}) + \sum_{j=1}^{p} |\beta_j / \tau s_j|^\alpha \right)$$



ouch.

To find the MAP, just iterate three steps

$$
\begin{aligned}
\beta^{(g+1)} &= \left(\tau^{-2} S^{-1} \hat{\Lambda}^{-1(g)} + \mathbf{X}_\star^T \hat{\Omega}^{-1(g)} \mathbf{X}_\star\right)^{-1} \left(\frac{1}{2} \mathbf{X}_\star^T \mathbf{1}\right) \\
\hat{\omega}_i^{-1(g+1)} &= \frac{1}{z_i^{(g)}} \left\{ \frac{e^{z_i^{(g)}}}{1 + e^{z_i^{(g)}}} - \frac{1}{2} \right\} \\
\hat{\lambda}_j^{-1(g+1)} &= \alpha (\tau s_j)^{2-\alpha} |\beta_j^{(g)}|^{\alpha-2}
\end{aligned}
$$

To find the MAP, just iterate three steps

Don't actually do this.

$$\beta^{(g+1)} = \left(\tau^{-2} S^{-1} \hat{\Lambda}^{-1(g)} + \mathbf{X}_\star^T \hat{\Omega}^{-1(g)} \mathbf{X}_\star\right)^{-1} \left(\frac{1}{2} \mathbf{X}_\star^T \mathbf{1}\right)$$

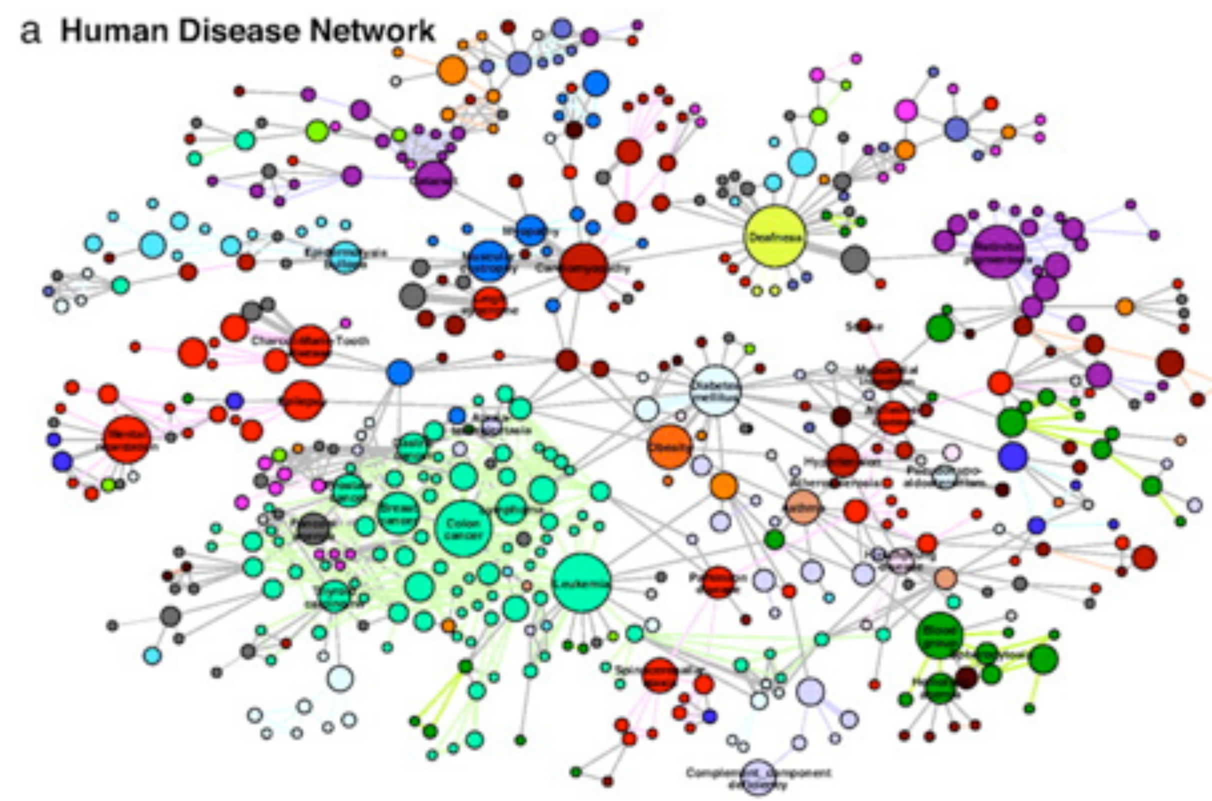
$$\hat{\omega}_i^{-1(g+1)} = \frac{1}{z_i^{(g)}} \left\{ \frac{e^{z_i^{(g)}}}{1 + e^{z_i^{(g)}}} - \frac{1}{2} \right\}$$

$$\hat{\lambda}_j^{-1(g+1)} = \alpha (\tau s_j)^{2-\alpha} |\beta_j^{(g)}|^{\alpha-2}$$

Example: covariate-dependent disease networks

~112m patient records comprising ICD-9 codes + prescriptions + covariates

ISCHEMIC HEART DISEASE (410-414)
410 Acute myocardial infarction
410.0 Of anterolateral wall
410.1 Of other anterior wall
410.2 Of inferolateral wall
410.3 Of inferoposterior wall
410.4 Of other inferior wall
410.5 Of other lateral wall
410.6 True posterior wall infarction
410.7 Subendocardial infarction
410.8 Of other specified sites
410.9 Unspecified site



Our approach: a tree of graphs

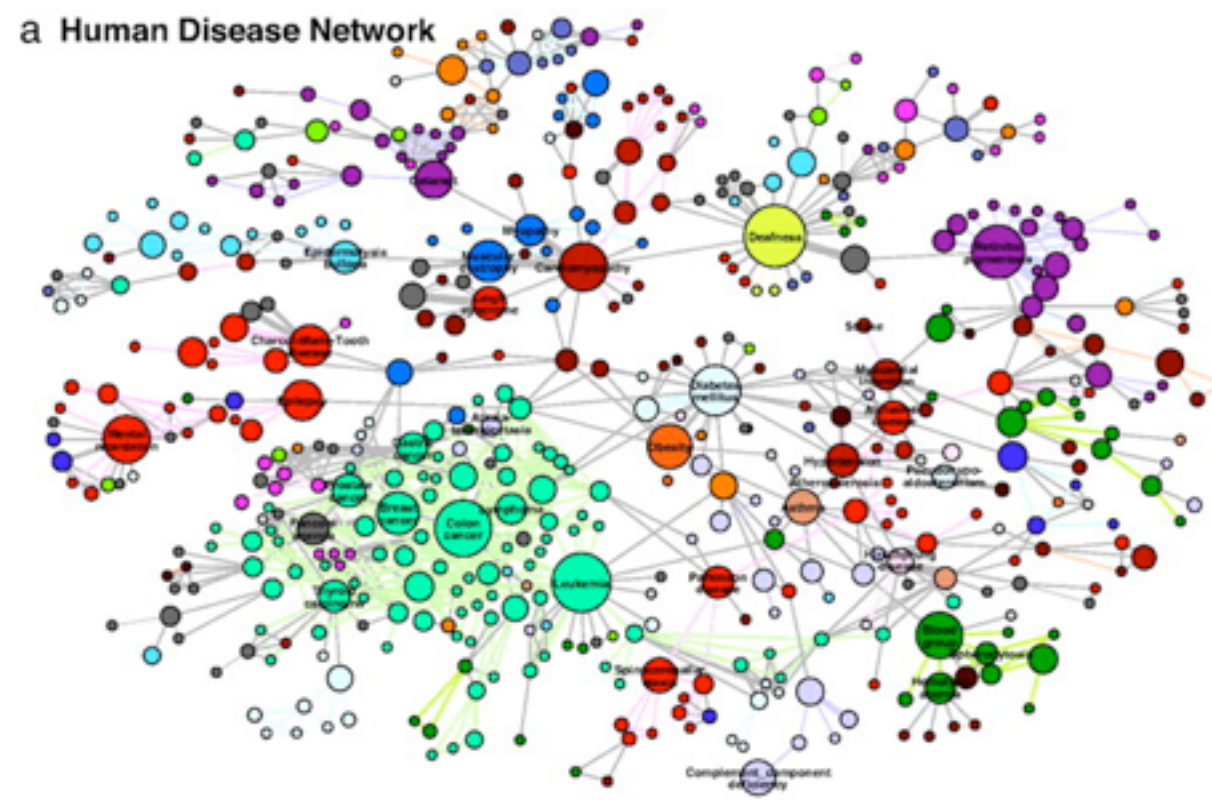
Tree splits on covariates, mainly demographics and geography.

Each terminal node is a disease network for a subgroup of the population.

Example: covariate-dependent disease networks

~112m patient records comprising ICD-9 codes + prescriptions + covariates

ISCHEMIC HEART DISEASE (410-414)
410 Acute myocardial infarction
410.0 Of anterolateral wall
410.1 Of other anterior wall
410.2 Of inferolateral wall
410.3 Of inferoposterior wall
410.4 Of other inferior wall
410.5 Of other lateral wall
410.6 True posterior wall infarction
410.7 Subendocardial infarction
410.8 Of other specified sites
410.9 Unspecified site



Goh et. al., PNAS (2007)

Our approach: a tree of graphs

Tree splits on covariates, mainly demographics and geography.

Each terminal node is a disease network for a subgroup of the population.

The big picture

We use normal variance–mean mixtures to represent a wide class of objective functions commonly encountered in high-dimensional problems.

By modern Bayesian standards, these are pretty simple.

But they are ubiquitous, useful, and often the only tractable approach in their respective domains for data sets beyond a certain size.

This class is surprisingly broad. For example:

$$p(\alpha_{1:K}, \beta_{1:N_d} \mid W, \mu, \Sigma) \propto \prod_{d=1}^{D} \left\{ p(\beta_d \mid \mu, \Sigma) \cdot \prod_{n=1}^{N_d} \left[\sum_{k=1}^{K} \left(\frac{e^{\beta_{dk}}}{\sum_{l=1}^{K} e^{\beta_{dl}}} \right) \alpha_{k,w_n} \right] \right\} \cdot p(\alpha_{1:K})$$

We get an exact MAP estimate.
No variational approximation.

This class is surprisingly broad. For example:

$$p(\alpha_{1:K}, \beta_{1:N_d} \mid W, \mu, \Sigma) \propto \prod_{d=1}^{D} \left\{ p(\beta_d \mid \mu, \Sigma) \cdot \prod_{n=1}^{N_d} \left[\sum_{k=1}^{K} \left(\frac{e^{\beta_{dk}}}{\sum_{l=1}^{K} e^{\beta_{dl}}} \right) \alpha_{k,w_n} \right] \right\} \cdot p(\alpha_{1:K})$$

(e.g. Blei and Lafferty, 2009)

We get an exact MAP estimate.
No variational approximation.

Connection with previous work

Penalties/priors corresponding to scale mixtures:

- lasso (Tibshirani, 1996; Park and Casella, 2008; Hans, 2009)
- bridge estimators (West, 1987; Huang et al., 2008)
- relevance vector machines (Tipping, 2001)
- normal/Jeffreys (Figueiredo, 2003; Bae and Mallick, 2004)
- normal/exponential-gamma (Griffin and Brown, 2005)
- normal/inverse-Gaussian (Caron and Doucet, 2008)
- normal/gamma (Griffin and Brown, 2010)
- horseshoe/inverted-beta prior (CPS 2010; Polson and Scott, 2011)
- double-Pareto (Armagan et al., 2010)

Algorithms for regularized regression

- LARS (Efron et al., 2004)
- LQA (Fan and Li, 2001) and LLA (Zou and Li, 2008)
- EM/ECME (Dempster et al., 1977; Meng and Rubin, 1993; many others)
- MM (Hunter and Lange, 2000; Taddy, 2010)
- MCMC for support-vector machines (Polson and Steve Scott, 2011)
- MCMC for logistic regression (Gramacy and Polson; Holmes and Held; Steve Scott; SFS; others)

Distributional theory based on variance-mean mixtures

- Z distributions (Fisher, 1923)
- Generalized inverse-Gaussian distributions (Barndorff-Nielsen 1977)
- Penalties, priors, and Lévy processes (Polson and Scott, 2011)

The standard scale-mixture trick

$$p(z) = \int_0^\infty \phi(z \mid v)\, p(v)\, dv$$

$$\hat{\beta} = \arg\min \left\{ ||\mathbf{y} - X\beta||^2 + \nu \sum_{j=1}^{p} g(\beta_j) \right\}$$

$$\begin{aligned}
(\mathbf{y} \mid \beta) &\sim \mathrm{N}(X\beta, \sigma^2 \beta) \\
(\beta_i \mid \tau^2, \lambda_i^2) &\sim \mathrm{N}(0, \tau^2 \lambda_i^2) \\
\lambda_i^2 &\sim \pi(\lambda_i^2) \\
(\tau^2, \sigma^2) &\sim \pi(\tau^2, \sigma^2)
\end{aligned}$$

$$\begin{aligned}
(\beta \mid \mathbf{y}, \tau^2, \Lambda, \sigma^2) &\sim \mathrm{N}(\hat{\beta}, \hat{\Sigma}_{\beta}) \\
\hat{\beta} &= (\tau^{-2}\Lambda^{-1} + X^T X)^{-1} X^T \mathbf{y}
\end{aligned}$$

$$\hat{\beta} = \arg\min \left\{ ||\mathbf{y} - X\beta||^2 + \nu \sum_{j=1}^{p} g(\beta_j) \right\}$$

$$\updownarrow$$

$$\begin{aligned}
(\mathbf{y} \mid \beta) &\sim \mathrm{N}(X\beta, \sigma^2\beta) \\
(\beta_i \mid \tau^2, \lambda_i^2) &\sim \mathrm{N}(0, \tau^2\lambda_i^2) \\
\lambda_i^2 &\sim \pi(\lambda_i^2) \\
(\tau^2, \sigma^2) &\sim \pi(\tau^2, \sigma^2)
\end{aligned}$$

Exponential —> Lasso
Inverted-beta —> Horseshoe
Gamma —> Normal/gamma
(Andrews and Mallows; West)

$$\begin{aligned}
(\beta \mid \mathbf{y}, \tau^2, \Lambda, \sigma^2) &\sim \mathrm{N}(\hat{\beta}, \hat{\Sigma}_\beta) \\
\hat{\beta} &= (\tau^{-2}\Lambda^{-1} + X^T X)^{-1} X^T \mathbf{y}
\end{aligned}$$

The standard scale-mixture trick

$$p(z) = \int_0^\infty \phi(z \mid v)\; p(v)\, dv$$

The standard scale-mixture trick

$$p(z) = \int_0^\infty \phi(z \mid v)\, p(v)\, dv$$

Two ways of generalizing this:

the Lévy representation

variance-mean mixtures

1) The Lévy representation

Let $\psi(t)$, $t > 0$, be a nonnegative-real-valued, totally monotone function such that $\lim_{t \to 0} \psi(t) = 0$.

Part A: Suppose that these conditions are met for $t \equiv f(\beta_j)$. Then the prior distribution $p(\beta_j \mid s) \propto \exp\{-s\psi[f(\beta_j)]\}$, where $s > 0$, is the moment-generating function of a subordinator $T(s)$, evaluated at $f(\beta_j)$, whose Lévy measure satisfies

$$\psi(t) = \int_0^\infty \{1 - \exp(-tx)\}\, \mu(\mathrm{d}x)\,. \tag{1}$$

Part B: Suppose that these conditions are met for $t \equiv \beta_j^2/2$. Then $p(\beta_j \mid s) \propto \exp\{-s\psi(\beta_j^2/2)\}$, where $s > 0$, is a mixture of normals given by

$$p(\beta_j \mid s) \propto \int_0^\infty \mathrm{N}\left(\beta_j \mid 0, T^{-1}\right)\, T^{-1/2} p(T)\, dT\,,$$

where $p(T)$ is the density of the subordinator T, observed at time s, whose Lévy measure $\mu(dx)$ satisfies (1).

1) The Lévy representation

Let $\psi(t)$, $t > 0$, be a nonnegative-real-valued, totally monotone function such that $\lim_{t \to 0} \psi(t) = 0$.

Part A: Suppose that these conditions are met for $t \equiv f(\beta_j)$. Then the prior distribution $p(\beta_j \mid s) \propto \exp\{-s\psi[f(\beta_j)]\}$, where $s > 0$, is the moment-generating function of a subordinator $T(s)$, evaluated at $f(\beta_j)$, whose Lévy measure satisfies

$$\psi(t) = \int_0^\infty \{1 - \exp(-tx)\}\, \mu(\mathrm{d}x)\,. \tag{1}$$

Part B: Suppose that these conditions are met for $t \equiv \beta_j^2/2$. Then $p(\beta_j \mid s) \propto \exp\{-s\psi(\beta_j^2/2)\}$, where $s > 0$, is a mixture of normals given by

$$p(\beta_j \mid s) \propto \int_0^\infty \mathrm{N}(\beta_j \mid 0, T^{-1})\; T^{-1/2} p(T)\, dT\,,$$

where $p(T)$ is the density of the subordinator T, observed at time s, whose Lévy measure $\mu(dx)$ satisfies (1).

(super useful for block-wise penalties, e.g. the group lasso)

2) Mean-variance mixtures

$$\cancel{p(z) = \int_0^\infty \phi(z \mid v)\, p(v)\, dv}$$

$$p(z) = \int_0^\infty \phi(z \mid \mu + kv, v)\, p(v)\, dv$$

(like Brownian motion with a drift; useful for non-Gaussian likelihoods)

The generic regularization problem

$$Q(\beta) = \sum_{i=1}^{n} f(y_i, x_i^T \beta) + \sum_{j=1}^{p} g\left(\frac{\beta_j}{\tau s_j}\right)$$

We consider properties of the posterior distribution

$$\begin{aligned} e^{-Q(\beta)} \propto p(\beta \mid \tau, y) &\propto \exp\left\{ -\sum_{i=1}^{n} f(y_i, x_i' \beta) - \sum_{j=1}^{p} \mathrm{g}(\beta_j / \tau s_j) \right\} \\ &\propto \left\{ \prod_{i=1}^{n} p(z_i \mid x_i^T \beta) \right\} \left\{ \prod_{j=1}^{k} p(\beta_j \mid \tau) \right\} \\ &= p(z \mid \beta) \cdot p(\beta \mid \tau), \end{aligned}$$

where $z_i = y_i - x_i^T \beta$ for regression, or $z_i = y_i x_i^T \beta$ for classification.

The generic regularization problem

Suppose that both the likelihood and prior/penalty can be represented as normal variance-mean mixtures.

$$
\begin{aligned}
p(z_i \mid \beta) &= \int_0^\infty \phi(z_i \mid \mu_z + \varkappa_z \omega_i, \sigma^2 \omega_i)\, dP(\omega_i) \\
p(\beta_j \mid \tau) &= \int_0^\infty \phi(\beta_j \mid \mu_\beta + \varkappa_\beta \lambda_j, \tau^2 s_j^2 \lambda_j)\, dP(\lambda_j)\,.
\end{aligned}
$$

Then we have an easy EM:

E Step

Compute the expected value of the log posterior, given the current parameter estimate:

$$Q(\beta \mid \beta^{(g)}) = \int \log p(\beta \mid \omega, \lambda, \tau, y) p(\omega, \lambda \mid \beta^{(g)}, \tau, y) \, d\omega \, d\lambda .$$

M Step

Maximize the complete-data posterior to update the parameter estimate:

$$\beta^{(g+1)} = \arg\max_{\beta} Q(\beta \mid \beta^{(g)}) .$$

The E Step

The observed-data log posterior is

$$p(\beta \mid \tau, y) = \int \pi(\beta \mid \omega, \lambda, y)\, p(\omega, \lambda \mid y, \tau)\, d\omega\, d\lambda .$$

Exploiting the mean–variance mixture representation, the complete-data log posterior is

$$\log p(\beta \mid \omega, \lambda, \tau, y) = c_0(\omega, \lambda, y, \tau) - \frac{1}{2} \sum_{i=1}^{n} \omega_i^{-1} \left(z_i - \mu_z - \varkappa_y \omega_i \right)^2 - \frac{1}{2 s_j^2 \tau^2} \sum_{j=1}^{p} \lambda_j^{-1} (\beta_j - \mu_\beta - \varkappa_\beta \lambda_j)^2$$

This can be shown to depend linearly upon the conditional moments $\{\hat{\omega}_i^{-1}\}$ and $\{\hat{\lambda}_j^{-1}\}$. Therefore the E-step is to simply plug these conditional expected values into the complete-data log posterior.

The E Step

Theorem: these conditional moments are:

$$(\beta_j - \mu_\beta)\hat{\lambda}_j^{-1(g)} = \varkappa_\beta + \tau^2 s_j^2\, g'(\beta_j \mid \tau),$$
$$(z_i - \mu_z)\hat{\omega}_i^{-1(g)} = \varkappa_z + \sigma^2\, f'(z_i \mid \beta),$$

Key fact: we don't need the conditional posterior for the latent variances.

Only need the functional form of the likelihood (f) and prior (g).

These are pre-specified.

Tilted, iteratively re-weighted least squares

E step

Given a current estimate $\beta = \beta^{(g)}$, compute the conditional moments of the latent variances as

$$\begin{aligned}
(\beta_j^{(g)} - \mu_\beta)\hat{\lambda}_j^{-1(g)} &= \kappa_\beta + \tau^2 s_j^2 \, g'(\beta_j^{(g)} \mid \tau), \\
(z_i^{(g)} - \mu_z)\hat{\omega}_i^{-1(g)} &= \kappa_z + \sigma^2 \, f'(z_i^{(g)} \mid \beta).
\end{aligned}$$

M Step

For regression, compute $\beta^{(g+1)}$ as

$$\beta^{(g+1)} = \left(\tau^{-2} S^{-1} \hat{\Lambda}^{-1(g)} + \mathbf{X}^T \hat{\Omega}^{-1(g)} \mathbf{X}\right)^{-1} \mathbf{X}^T \left(\hat{\Omega}^{-1(g)} y - \mu_z \omega^{-1(g)} - \kappa_z \mathbf{1}\right).$$

For classification, compute $\beta^{(g+1)}$ as

$$\beta^{(g+1)} = \left(\tau^{-2} S^{-1} \hat{\Lambda}^{-1(g)} + \mathbf{X}_\star^T \hat{\Omega}^{-1(g)} \mathbf{X}_\star\right)^{-1} \mathbf{X}_\star^T \hat{\Omega}^{-1(g)} \left(\mu_z \mathbf{1} + \kappa_z \hat{\omega}^{(g)}\right).$$

This works for a broad class of models.

Likelihood	$f(z_i \mid \beta)$	k_z	μ_z	$p(\omega_i)$
Squared-error	z_i^2	0	0	$\omega_i \equiv 1$
Absolute-error	$\lvert z_i \rvert$	0	0	Exponential
Check loss	$\lvert z_i \rvert + (2q-1)z_i$	$1-2q$	0	GIG
SVM	$\max(1-z_i, 0)$	1	1	GIG
Logistic	$\log(1+e^{z_i})$	$1/2$	0	Polya

Penalty function	$g(\beta_j \mid \tau)$	k_β	μ_β	$p(\lambda_j)$
Ridge	$(\beta_j/\tau)^2$	0	0	$\omega_i \equiv 1$
Lasso	$\lvert \beta_j/\tau \rvert$	0	0	Exponential
Bridge	$\lvert \beta_j/\tau \rvert^\alpha$	0	0	Stable
Gen. Double-Pareto	$\{(1+\alpha)/\tau\} \log(1+\lvert \beta_j \rvert/\alpha\tau)$	0	0	Exp–Gam

+ multinomial logit, autologistic, topic models, RBMs, extreme-value ....

Two key identities:

$$\frac{\alpha^2 - \varkappa^2}{2\alpha} e^{-\alpha|\theta-\mu|+\varkappa(\theta-\mu)} = \int_0^\infty \phi(\theta \mid \mu + \varkappa v, v))\; p_{gig}(v \mid 1, 0, \sqrt{\alpha^2 - \varkappa^2})\; dv$$

$$\frac{1}{B(\alpha,\varkappa)} \frac{e^{\alpha(\theta-\mu)}}{(1+e^{\theta-\mu})^{2(\alpha-\varkappa)}} = \int_0^\infty \phi(\theta \mid \mu + \varkappa v, v)\; p_{pol}(v \mid \alpha, \alpha - 2\varkappa)\, dv\,.$$

Improper versions (treat purely as an integral identity)

$$a^{-1} \exp\left\{-2c^{-1} \max(au, 0)\right\} = \int_0^\infty \phi(u \mid -av, cv)\, dv$$

$$c^{-1} \exp\left\{-2c^{-1} \rho_q(u)\right\} = \int_0^\infty \phi(u \mid -(2\tau - 1)v, cv) e^{-2\tau(1-\tau)v}\, dv$$

$$(1 + \exp\{u - \mu\})^{-1} = \int_0^\infty \phi(u \mid \mu - (1/2)v, v)\; p_{pol}(v \mid 0, 1)\, dv$$

where $\rho_q(u) = \frac{1}{2}|u| + \left(q - \frac{1}{2}\right) u$ is the check-loss function.

Back to the teaser example: logit models

A Polya mixing distribution ...

$$p_{pol}(v \mid \alpha, \alpha - 2k) = \sum_{k=0}^{\infty} w_k e^{-a_k v}$$

The terms in this sum are

$$\begin{aligned} a_k &= \frac{(\alpha + k)(k + k)}{2} \\ w_k &= a_k \prod_{j \neq k} \left(\frac{a_k}{a_j - a_k} \right) = \left(\begin{array}{c} -2\delta \\ k \end{array} \right) \frac{(\delta + k)}{B(\delta + b, \delta - b)}, \end{aligned}$$

where $b = \frac{1}{2}(\alpha - k)$, $\delta = \frac{1}{2}(\alpha + k)$, and $\left(\begin{array}{c} -2\delta \\ k \end{array} \right) = \frac{(-1)^k (2\delta)...(2\delta + k - 1)}{k!}$.

Back to the teaser example: logit models

A Polya mixing distribution leads to a Z distribution marginal.

$$p_Z(\theta \mid \mu, \alpha, k) = \frac{1}{B(\alpha, k)} \frac{(e^{\theta - \mu})^{\alpha}}{(1 + e^{\theta - \mu})^{2(\alpha - k)}}$$

This family includes the logistic regression model as a limiting, improper case: (a, k, μ) = (1, 1/2, 0).

Our representation theorem gives the relevant conditional moment as

$$\hat{\omega}_i^{-1} = \frac{1}{z_i} \left\{ \frac{e^{z_i}}{1 + e_i^{z}} - \frac{1}{2} \right\}, \; z_i = y_i x_i^T \beta$$

Multinomial logit

The probability that observation i falls into class k is

$$\theta_{ki} = P(y_i = k) = \frac{\exp(x_i^T \beta_k)}{\sum_{l=1}^{K} \exp(x_i^T \beta_l)}$$

To represent the multinomial logit model in our framework, let

$$\begin{aligned} \eta_{ki} &= \exp\left(x_i^T \beta_k - c_{ki}\right) / \{1 + \exp\left(x_i^T \beta_k - c_{ki}\right)\} \\ c_{ki}(\beta_{(-k)}) &= \log \sum_{l \neq k} \exp\{x_i^T \beta_l\} \end{aligned}$$

(Holmes and Held, 2006)

Multinomial logit

Write the conditional likelihood for category k as

$$
\begin{aligned}
L(\beta_k \mid \beta_{(-k)}, y) &\propto \prod_{i=1}^{n} \prod_{l=1}^{K} \theta_{li}^{\tilde{y}_{li}} \\
&\propto \prod_{i=1}^{n} \eta_{ki}^{\tilde{y}_{ki}} \{w_i(1-\eta_{ki})\}^{1-\tilde{y}_{ki}} \\
&\propto \prod_{i=1}^{n} \eta_{ki}^{\tilde{y}_{ki}} \{(1-\eta_{ki})\}^{1-\tilde{y}_{ki}} \\
&\propto \prod_{i=1}^{n} \left\{ \frac{\exp(\gamma_{ki} x_i^T \beta_k - \gamma_{ki} c_{ki})}{1+\exp(\gamma_{ki} x_i^T \beta_k - \gamma_{ki} c_{ki})} \right\} \\
&= \prod_{i=1}^{n} \int_0^\infty \phi(z_{ki} \mid \mu_{ki} + \varkappa \xi_{ki}, \xi_{ki})\, p_{PY}(\xi_{ki} \mid 1, 0)\, d\xi_{ki}
\end{aligned}
$$

where $\varkappa = 1/2$, $z_{ki} = \gamma_{ki} x_i^T \beta_k$, $\mu_{ki} = \gamma_{ki} c_{ki}$, and $p_{PY}(\xi_{ki} \mid 1, 0)$ is a function of ξ_{ki} that is the limit of a Polya density as $b \to 0$.

An exact ECM

For iteration t = 1, 2, . . .

For $k = 2, \ldots, K$, cycle through the following steps:

Update Ω_k**:**

$$
\begin{aligned}
z_{ki}^{(t)} &:= \gamma_{ki} x_i^T \beta_k^{(t)} \\
\mu_{ki}^{(t)} &:= \gamma_{ki} \log \sum_{l \neq k} \exp(x_i^T \beta_l^{(t)}) \\
\omega_{ki}^{(t)} &:= \left(\frac{1}{z_{ki}^{(t)} - \mu_{ki}^{(t)}} \right) \left(\frac{\exp\left\{ z_{ki}^{(t)} - \mu_{ki}^{(t)} \right\}}{1 + \exp\left\{ z_{ki}^{(t)} - \mu_{ki}^{(t)} \right\}} - \frac{1}{2} \right) \\
\Omega_k^{(t)} &:= \mathrm{diag}(\omega_{k1}^{(t)}, \ldots, \omega_{kn}^{(t)}),
\end{aligned}
$$

where $\beta_k^{(t)}$ is the current estimate for the kth block of coefficients, and where $\gamma_{ki} = \pm 1$ is an indicator of whether $y_i = k$.

An exact ECM

For iteration t = 1, 2, . . .

For $k = 2, \ldots, K$, cycle through the following steps:

Update Λ_k**:**

$$\begin{aligned} \lambda_{kj} &:= \frac{\tau^2 \psi'(\beta_{kj}^{(t)}/\tau)}{\beta_{kj}^{(t)}} \\ \Lambda_k^{(t)} &:= \mathrm{diag}(\lambda_{k1}^{(t)}, \ldots, \lambda_{kp}^{(t)}), \end{aligned}$$

where ψ' is the derivative of the penalty function $\psi(\beta_{kj})$.

An exact ECM

For iteration t = 1, 2, . . .

For $k = 2, \ldots, K$, cycle through the following steps:

Update β_k**:** Solve the linear system $A_k^{(t)} \beta_k^{(t+1)} = b_k^{(t)}$ for $\beta_k^{(t+1)}$, with

$$\begin{aligned} A_k^{(t)} &:= \tau^{-2} \Lambda_k^{(t)} + \tilde{X}_k^T \Omega_k^{(t)} \tilde{X}_k \\ b_k^{(t)} &:= \tilde{X}_k^T \left(\Omega_k^{(t)} \mu_k^{(t)} + \frac{1}{2} \mathbf{1} \right), \end{aligned}$$

where $\tilde{X}_k$ is the $n \times p$ matrix having rows $\tilde{\mathbf{x}}_i = \gamma_{ki} x_i$, $\mathbf{1}$ is a column vector of ones, and $\mu_k^{(t)} = (\mu_{k1}^{(t)}, \ldots, \mu_{kn}^{(t)})^T$.

An exact ECM

For iteration t = 1, 2, ...

For $k = 2, \ldots, K$, cycle through the following steps:

Update β_k**:** Solve the linear system $A_k^{(t)} \beta_k^{(t+1)} = b_k^{(t)}$ for $\beta_k^{(t+1)}$, with

$$\begin{aligned} A_k^{(t)} &:= \tau^{-2} \Lambda_k^{(t)} + \tilde{X}_k^T \Omega_k^{(t)} \tilde{X}_k \\ b_k^{(t)} &:= \tilde{X}_k^T \left(\Omega_k^{(t)} \mu_k^{(t)} + \frac{1}{2} \mathbf{1} \right), \end{aligned}$$

where $\tilde{X}_k$ is the $n \times p$ matrix having rows $\tilde{\mathbf{x}}_i = \gamma_{ki} x_i$, $\mathbf{1}$ is a column vector of ones, and $\mu_k^{(t)} = (\mu_{k1}^{(t)}, \ldots, \mu_{kn}^{(t)})^T$.

Don't solve this exactly.

Tilted, iteratively re-weighted conjugate gradient

In the update step for ß, don't solve the system exactly. Instead:

While $|\Delta_{(t,l)}| > \delta_{\min}$, increment l and set

$$
\begin{aligned}
\beta^{(c,l)} &:= \beta_k^{(c,l-1)} + \Delta_{(t,l-1)} \\
r_{(t,l)} &:= r_{(t,l-1)} - \alpha_{(t,l-1)} c_{(t,l-1)} \\
\gamma_{(t,l)} &:= \frac{r_{(t,l)}^T r_{(t,l)}}{r_{(t,l-1)}^T r_{(t,l-1)}} \\
d_{(t,l)} &:= r_{(t,l)} + \gamma_{(t,l)} d_{(t,l-1)} \\
c_{(t,l)} &:= A_k^{(t)} d_{(t,l)} \\
\alpha_{(t,l)} &:= \frac{r_{(t,0)}^T r_{(t,l)}}{d_{(t,l)}^T c_{(t,l)}} \\
\Delta_{(t,l)} &:= \alpha_{(t,l)} d_{(t,l)}.
\end{aligned}
$$

Tilted, iteratively re-weighted conjugate gradient

In the update step for ß, don't solve the system exactly. Instead:

While $|\Delta_{(t,l)}| > \delta_{\min}$, increment l and set

$$
\begin{aligned}
\beta^{(c,l)} &:= \beta_k^{(c,l-1)} + \Delta_{(t,l-1)} \\
r_{(t,l)} &:= r_{(t,l-1)} - \alpha_{(t,l-1)} c_{(t,l-1)} \\
\gamma_{(t,l)} &:= \frac{r_{(t,l)}^T r_{(t,l)}}{r_{(t,l-1)}^T r_{(t,l-1)}} \\
d_{(t,l)} &:= r_{(t,l)} + \gamma_{(t,l)} d_{(t,l-1)} \\
c_{(t,l)} &:= A_k^{(t)} d_{(t,l)} \\
\alpha_{(t,l)} &:= \frac{r_{(t,0)}^T r_{(t,l)}}{d_{(t,l)}^T c_{(t,l)}} \\
\Delta_{(t,l)} &:= \alpha_{(t,l)} d_{(t,l)}.
\end{aligned}
$$

Parallelize this.

Other methods

You can solve many of these problems using tailored methods—but not all of them, and rarely this simply or efficiently.

LARS: efficient only in Gaussian models

Coordinate descent: can get stuck; may require numerical derivatives; irredeemably serial

LLA/LQA: exact only in Gaussian models (where it's a special case of our method)

Variational methods: never exact; sometimes inconsistent; poor in cases of interesting a posteriori dependence

How would you do MCMC in this class?

Some people have worked out special cases.

Here's an interesting fact:

$$
\begin{aligned}
p_{pol\left(\frac{1}{2},\frac{1}{2}\right)}(\lambda) &= \sum_{n=0}^{\infty}(-1)^n\left(n+\frac{1}{2}\right)e^{-\frac{1}{2}\left(n+\frac{1}{2}\right)^2\lambda} \\
C &\stackrel{D}{=} \frac{1}{4\pi^2}\lambda
\end{aligned}
$$

$$
\begin{aligned}
p(C) &= 4\pi^2\sum_{n=0}^{\infty}(-1)^n\left(n+\frac{1}{2}\right)e^{-2\left(n+\frac{1}{2}\right)^2\pi^2 C} \\
C &\stackrel{D}{=} \frac{1}{2\pi^2}\sum_{n=1}^{\infty}\frac{\mathcal{E}(1)}{\left(n-\frac{1}{2}\right)^2}
\end{aligned}
$$

How would you do MCMC in this class?

There is a duality between the scale-mixture and MGF representations for C:

$$\mathbb{E}\left(e^{-\frac{x^2}{2}C}\right) = \mathbb{E}\left((2\pi C)^{-\frac{1}{2}} e^{-\frac{x^2}{8\pi^2 C}}\right) = \frac{1}{cosh\left(\frac{x}{2}\right)}$$

A logit-type likelihood would thus look like

$$\begin{aligned} \frac{2^{a+b} e^{ax}}{(1+e^x)^{a+b}} &= e^{\frac{1}{2}(a-b)x} \cdot \left(\frac{1}{cosh\left(\frac{x}{2}\right)}\right)^{a+b} \\ &= e^{\frac{1}{2}(a-b)x} \cdot \mathbb{E}\left(e^{-\frac{x^2}{2}C}\right) \end{aligned}$$

How would you do MCMC in this class?

Therefore we can simulate from C using Polya-Gamma distributions:

$$
\begin{aligned}
(C \mid x) &\overset{D}{=} 2\sum_{n=1}^{\infty} \frac{\mathcal{G}(a+b,1)}{a_n} \\
a_n &= \frac{1 + \frac{x^2}{4\left(n-\frac{1}{2}\right)^2 \pi^2}}{4\left(n-\frac{1}{2}\right)^2 \pi^2}
\end{aligned}
$$

Robust regression	Student t
Logit models	Ridge
Multinomial logit models	Lasso
Extreme-value models	Bridge
Support vector machines	Normal/exponential-gamma
Topic models	Normal/gamma
Restricted Boltzmann machines	Generalized-t (double Pareto)
Neural networks	Normal/inverted-beta
Autologistic models	Normal/inverse-Gaussian
Penalized additive models	Meridian filter

The theory of variance-mean mixtures allows MAP estimation for arbitrary combinations of model (left) with prior or penalty (right).

With the conjugate-gradient version, there are no matrix inverses and no numerical derivatives.

Three papers

"Sparse Bayes estimation in non-Gaussian models via data augmentation."

"Sparse multinomial logistic regression via nonconcave penalized likelihood."

"Exact MAP estimation in logistic-normal topic models."