Two Applications of Variational Inference: Astrophysics and Gene Expression Analysis

Ata Kaban
School of Computer Science
The University of Birmingham
Overview

• Part I:
  – Robust mixtures for data with error information: An astrophysical application.
• Part II:
  – On Bayesian classification with Laplace priors: An application to micro-arrays.
Robust mixtures in the presence of measurement errors

- Robust unsupervised modelling of data aims at capturing the structure of the typical observations while dealing with the atypical ones automatically.
- Dealing with outliers is required to avoid biases of parameter estimates and poor generalisation of the structure inferred from the data.
- In addition, detecting peculiar objects from large data repositories is of special interest e.g. in Astrophysics, where it provides a unique means of identifying candidates of possibly new types of objects (e.g. from archives of multi-wavelength astronomical images) that deserve follow-up study (e.g. using spectroscopy).
• Bottleneck: ‘the likely overabundance of interesting objects found’ – the interpretation and understanding of which will necessitate costly detailed analysis

• Indeed, not every atypical observation is truly interesting
• One reason for this lies in that measurement errors can produce outliers.
• These errors are due to instrument calibration, physical limitations of devices and experimental conditions.
• These errors are recorded and are available
• Yet, most of existing data analysis / data mining methods have no natural ways of taking these into consideration.
• In turn, neglecting these errors holds the risk of compromising our ability of detecting the ‘interesting’ outliers
• ‘Errors in variables’ approaches exist for robust regression and probabilistic approaches able to propagate uncertainty are starting to appear for certain problems
• We extend this spectrum for unsupervised robust density modelling
Finite Mixture Models (for Clustering)

Generative process:
- Sample a class $z_n = k \sim \pi$
- Sample a data point $y_n \sim p(y_n | \theta_k)$

Likelihood: $L(\theta) = p(y_n | \theta) = \sum_k p(y_n | z_n = k, \theta_k) p(z_n = k)$

More generally...

“For many applied problems the tails of the normal distribution are shorter than required”

The Student-t distribution

The $t$ distributions were discovered by William Gosset in 1908. He wrote under the name ‘Student’

$$S(t) = \frac{\Gamma((k + 1)/2)}{\sqrt{k \pi \Gamma(k/2)}} (1 + t^2/k)^{-(k+1)/2}$$
The Multivariate Student-t distribution

\[ S_i(t | \mu, \Sigma, \nu) = \frac{\Gamma(\nu/2 + \text{dim}/2) |\Sigma|^{-1/2}}{\left(\pi \nu \right)^{\text{dim}/2} \Gamma(\nu/2)\left[1 + (t - \mu)^T \Sigma^{-1} (t - \mu) / \nu \right]^{(\nu/2 + \text{dim}/2)}} \]

A useful property: The t-distribution is a scale-mixture.
(Andrews & Mallows, 1974, J R Statist Soc B)

\[ = \int_0^\infty N(t | \mu, \Sigma / u, \nu) G(u | \nu / 2, \nu / 2) du \]

where

\[ N(t | \mu, \Sigma / u) = (2\pi)^{-\text{dim}/2} |\Sigma / u|^{-1/2} \exp[-1/2(t - \mu)^T (\Sigma / u)^{-1} (t - \mu)] \]

\[ G(u | a, b) = b^a u^{a-1} \exp(-bu) / \Gamma(a) \]
\[= \int_{0}^{\infty} N(t | \mu, \Sigma /u,v)G(u | \nu/2, \nu/2)du\]

- So the t-distribution can be thought of, and estimated as a latent variable model with a univariate Gamma latent variable.

- Graphical model equivalent to a multivariate t-distribution:

  Posterior expectation \(E[u|t_n]\) infers the ‘outlierness’ (peculiarity)

In this example we only have a single non-outlying high-density region. In order to allow for more, we need a mixture of t-distributions.
Mixture of t-distributions (Robust mixture)

- Two sets of latent variables:
  - A discrete class-variable \( z \)
  - A Gamma variable \( u \)
- Maximum likelihood estimation via EM
- Inferred ‘peculiarity’ = \( E[u|t_n] \), which is now obtained via marginalising over the class-variable.
Robust mixture in the presence of measurement errors

• Now the data is only observed up to some errors (square of errors are in the diagonals of $S_n$)

• Apart from clustering the data, we try to find the peculiar objects whose peculiarity is not due to variations caused by measurement errors (‘genuine outliers’)

• Exact maximum likelihood is no longer tractable to compute

• We developed
  - variational EM
  - MarkovChain-EM using Gibbs sampling
Robust mixtures for data with errors

- Data from $N$ objects, $d$ features, with elements of the form
  \[ t_{in} \pm \sqrt{s_{in}} \]

- Hetero-schedastic Gaussian error model is justified
  \[ p(t_n | w_n) = \mathcal{N}(t_n | w_n, S_n) \]

- We model the hidden clean density by a mixture of $t$.
  \[ p(w_n) = \sum_{k=1}^{K} \pi_k S_t(w_n | \mu_k, \Sigma_k, \nu_k) \]

- So the marginal data likelihood is:
  \[ p(t) = \sum_k \pi_k \int \int \mathcal{N}(t | w, S) \mathcal{N}\left(w | \mu_k, \frac{\Sigma_k}{u}\right) \mathcal{G}\left(u | \frac{\nu_k}{2}, \frac{\nu_k}{2}\right) du dw \]
• The joint likelihood of all variables:

\[
\prod_n \prod_k [p(t_n | w_n)p(w_n | u_n, z_n = k)] \delta(z_n = k) \times \\
\prod_n \prod_k [p(u_n | z_n = k)p(z_n = k | \pi)] \delta(z_n = k)
\]
Solving by variational EM

- Lower bounding the log likelihood – for any $q()$, we have:

$$\log p(t_n|\theta) = \int q(h_n) \log \frac{p(h_n, t_n|\theta)}{q(h_n)} \frac{q(h_n)}{p(h_n|t_n, \theta)} dh_n$$

$$\geq \int q(h_n) \log \frac{p(h_n, t_n|\theta)}{q(h_n)} dh \equiv \mathcal{F}(t_n|q, \theta)$$

where

$h_n = (w_n, u_n, z_n)$ the hidden variables in the model

$\theta = (\{\mu_k\}, \{\Sigma_k\}, \{v_k\}, \pi)$ the model parameters

- E-M with variational E-step at (k+1)-th iteration:
  - Var E step:
    $$q^{k+1}(h_n) = \arg \max_q \mathcal{F}(t_n|q, \theta^k)$$
  - M-step
    $$\theta^{k+1} = \arg \max_{\theta} \sum_n \mathcal{F}(t_n|q^{k+1}, \theta)$$
Choice of the variational distribution

- Option 1: Fully factorial form

\[ q(w, u, z) \equiv q(w)q(u)q(z) \]

- Option 2: Tree-structured factorial form

\[ q(w, u, z) \equiv \prod_{k=1}^{K} [q(w \mid z = k)q(u \mid z = k)q(z = k)] \delta(z=k) \]

In both cases, the optimal functional form of the terms of \( q() \) is tractable to compute, by taking the functional derivatives and equating them to the identically null function.

Schematic details for option 2 on the next few slides.
• Notations

denote $q(z = k)$ by $q(k)$, $q(w|z = k)$ by $q(w|k)$ and $q(u|z = k)$ by $q(u|k)$. Also, expectations w.r.t. $q(w|z = k)$ will be denoted by $\langle . \rangle_{w|k}$ and similarly, those w.r.t. $q(u|k)$ by $\langle . \rangle_{u|k}$, and those w.r.t. the joint $q(w,u|k) = q(w|k)q(u|k)$ by $\langle . \rangle_{w,u|k}$. 
The Var-E-step expressions

\[ q(w|k) = \frac{\exp\{\log [p(t|w)p(w|u, k)]_w|k\}}{\int \exp\{\log [p(t|w)p(w|u, k)]_w|k\} dw} = \mathcal{N}(w|\langle w \rangle_k, \Sigma_w|k) \]

\[ q(u|k) = \frac{\exp\{\log [p(w|u, k)p(u|k)]_w|k\}}{\int \exp\{\log [p(w|u, k)p(u|k)]_w|k\} du} = \mathcal{G}(u|a_k, b_k) \]

\[ q(k) = \frac{\exp(A_{t,k})}{\sum_{k'} \exp(A_{t,k'})} \]

where

\[ \Sigma_{w|k} = S \left[ \frac{\Sigma_k}{\langle u \rangle_u|k} + S \right]^{-1} \frac{\Sigma_k}{\langle u \rangle_u|k} \]

\[ \langle w \rangle_k = \Sigma_{w|k} (\langle u \rangle_u|k \Sigma_k^{-1} \mu_k + S^{-1} t) \]

\[ a_k = \frac{\nu_k + d}{2}; \quad b_k = \frac{\nu_k + C_k}{2} \]

\[ C_k = (\langle w \rangle_k - \mu_k)^T \Sigma_k^{-1} ((\langle w \rangle_k - \mu_k) + \text{Tr} (\Sigma_k^{-1} \Sigma_{w|k}) \]

and the expression \( A_{t,k} \) depends on all other variables and will be available from the likelihood bound computations.
The expression denoted by $A_{t,k}$ earlier is the following:

$$A_{t,k} = \langle \log p(t|w) \rangle_{w|k} + \langle \log p(u|k) \rangle_{u|k} + \log \pi_k + \langle \log p(w|u,k) \rangle_{w,u|k} - \langle \log q(w|k) \rangle_{w|k} - \langle \log q(u|k) \rangle_{u|k}$$

$$= Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6$$

$$Q_1 = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |S| - \frac{1}{2} \text{Tr}(\Sigma_{w|k} S^{-1}) - \frac{1}{2} \left[ (\langle w \rangle_k - t)^T S^{-1} (\langle w \rangle_k - t) \right]$$

$$Q_2 = \left( \frac{\nu_k}{2} - 1 \right) \langle \log u \rangle_{u|k} - \frac{\nu_k}{2} \frac{a_k}{b_k} + \frac{\nu_k}{2} \log \left( \frac{\nu_k}{2} \right) - \log \Gamma \left( \frac{\nu_k}{2} \right)$$

$$Q_3 = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_k| + \frac{d}{2} \frac{a_k}{b_k}$$

$$- \frac{1}{2} \frac{a_k}{b_k} \left[ (\langle w \rangle_k - \mu_k)^T \Sigma_k^{-1} (\langle w \rangle_k - \mu_k) - \frac{1}{2} \frac{a_k}{b_k} \text{Tr}(\Sigma_{w|k} \Sigma_k^{-1}) \right]$$

$$Q_4 = \log \pi_k;$$

$$Q_5 = \frac{d}{2} + \frac{d}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_{w|k}|$$

$$Q_6 = -[(a_k - 1) \langle \log u \rangle_{u|k} + a_k \log b_k - a_k - \log \Gamma(a_k)]$$

and where $\langle \log u \rangle_{u|k} = \psi(a_k) - \log b_k$ and $\psi(\cdot)$ is the di-gamma function.
The variational bound

- With the previous expression, then the likelihood bound for a set of $N$ data points is computed as the following:

$$
\mathcal{F} = \sum_{n} \sum_{k} q(z_n = k) \left[ A_{t_n,k} - \log q(z_n = k) \right]
$$

- At each iteration of the variational E-step and the M-step updates, the variational log likelihood bound is guaranteed not to decrease, and alternating these two steps will converge to a local optimum of the bound.

- Monitoring the likelihood bound is useful for assessing convergence of the iterative algorithm.
The M-step expressions

- Solving the stationary equations w.r.t. the model parameters yields:

\[
\mu_k = \frac{\sum_{n=1}^{N} q(z_n = k) \langle u_n \rangle_{u_n | k} \langle w_n \rangle_{w_n | k}}{\sum_{n=1}^{N} q(z_n = k) \langle u_n \rangle_{u_n | k}}
\]

\[
\Sigma_k = \frac{\sum_{n=1}^{N} q(z_n = k) \langle u_n \rangle_{u_n | k} \Sigma_{n,k}}{\sum_{n=1}^{N} q(z_n = k)}
\]

\[
\pi_k = \frac{1}{N} \sum_{n=1}^{N} q(z_n = k)
\]

where \( \Sigma_{n,k} = [(\mu_k - \langle w_n \rangle_k)(\mu_k - \langle w_n \rangle_k)^T + \Sigma_{w_n | k}] \).

\[
\sum_{n} q(z_n = k)[\log(\frac{\nu_k}{2}) + 1 + \langle \log u_n \rangle_k - \frac{a_{nk}}{b_{nk}} - \psi(\frac{\nu_k}{2})] = 0
\]
Computational scaling

- The time complexity per iteration is $O(d^3KN)$.
- For the fully factorial posterior choice, the theoretical complexity of the resulting algorithm, per iteration, would have been the same. The CPU time to convergence, evaluated empirically, was observed to take about 1.5 times longer in the case of the structured posterior.
- The time complexity of the ML estimation algorithm of t-mixtures (Peel & McLachlan), per iteration, is also $O(d^3KN)$.
- The only extra burden resulted from having taken measurement errors into account is the computation of the posterior statistics of the ‘clean’ latent variable $w$.
- The most expensive operation is the $d \times d$ matrix inversion. Note, $d$ is expected to be small, since in high dimensions $\Sigma_k$ would need to be assumed diagonal, in which the cubic operation disappears.
- Converge speed-ups are subject of future work.
Determining the number of components

- We used a lower bound to the Minimum Message Length in conjunction with our likelihood bound. Other methods are possible.
- The MML criterion is to maximise:

\[
\mathcal{L}(\theta, \mathcal{Y}) = -\frac{\hat{n}}{2} \sum_{k: \pi_k > 0} \log \left( \frac{N \pi_k}{12} \right) - \frac{k_{nz}}{2} \log \left( \frac{N}{12} \right) - \frac{k_{nz} (\hat{n} + 1)}{2} + \log p(\mathcal{Y}|\theta)
\]

where the last term is the data log likelihood – which we replace by our variational lower bound

\[
\sum_n \mathcal{F}(\mathbf{t}_n | q, \theta)
\]

and where

\(\hat{n}\) = dimensionality of the free parameters (e.g. in our case, \(\{\mathbf{\mu}_k, \Sigma_k\}\) have \(d + d(d-1)/2\) free params)

\(k_{nz}\) = # nonzero mixture components

This maximisation task is similar to that presented before, except the update of the mixture priors, which is now:

\[
\pi_k = \frac{\max \left\{0, \sum_{n=1}^{N} q(z_n = k) - \frac{\hat{n}}{2}\right\}}{\sum_{j=1}^{K} \max \left\{0, \sum_{n=1}^{N} q(z_n = j) - \frac{\hat{n}}{2}\right\}}
\]
Inferring the ‘interesting’ outliers

- The quantity of interest is the posterior expectation of \( u_n \). We use the variational expectation, computed wrt \( \sum q(u|k)q(k) \). The lower this expression is, the most outlying the associated observation. This expression is computed as the following:

\[
e \equiv \sum_{k} q(k) \frac{\nu_k + d}{\nu_k + \text{Tr}(\Sigma_k^{-1} \Sigma_w|k) + \Delta_{w|k}^2}
\]

where

\[
\Delta_{w|k}^2 = (\langle w \rangle_w|k - \mu_k)^T \Sigma_k^{-1} (\langle w \rangle_w|k - \mu_k)
\]

- For t-mixtures (i.e. no measurement error info), in turn, the posterior mean expression for \( u_n \) is:

\[
e_{\text{MoT}} \equiv \sum_{k} p(k|t) \frac{\nu_k + d}{\nu_k + (t - \mu_k)^T \Sigma_k^{-1} (t - \mu_k)}
\]

- It can easily be shown that in the limit of zero measurement errors, our variational posterior expressions recover the exact posteriors of t-mixtures (and the whole algorithm recovers the EM-based ML estimation algorithm of t-mixtures.)

- So, knowledge about the measurement errors re-defines the notion of ‘interesting’ outlierness.
gain more insights and see the effects of a misspecification of the error by rewriting the data likelihood by integrating over \( w \):

\[
p(t_n) = \sum_k \pi_k \int \mathcal{N} \left( t_n | \mu_k, \frac{\Sigma_k}{\nu_n} + S_n \right) \mathcal{G} \left( u | \frac{\nu_k}{2}, \frac{\nu_k}{2} \right) du_n
\]

The posterior expectations \( \langle u_n \rangle \) are data instance-specific, ensuring the robustness of the parameter estimates, even if the errors (diagonals of \( S_n \)) are misspecified. However, this also implies that a data instance with an underestimated \( S_n \) gets picked as a false ‘interesting’ outlier (\( \langle u_n \rangle \) gets smaller). Clearly, if all errors are specified at zero, our model reduces to MoT and produces unwanted false detections.
Experiments and Results

1) How accurate is the structured variational approximation employed?
   - How it compares to a fully factorial approximation?
   - How it compares to the ‘ground truth’ Markov-Chain-EM?

2) To what extent does knowledge of measurement errors help us to infer the outliers that are not due to these errors?
   - How it compares to ignoring the knowledge about the errors?
   - How it compares to the idealised case of having the clean data with no measurement errors?

3) Application in Astronomy: Finding peculiar objects of interest from the SDSS quasar catalogue (high-redshift quasars).
Illustration

- Synthetic data sampled from three well separated Gaussian and genuine outliers from a uniform distribution
- Gaussian noise added to each point to simulate the presence of measurement errors
- The aim is to recover the outliers despite the noise added

Left: Hidden error-free data with clusters and genuine outliers.
Centre: Data contaminated with measurement errors
Right: The estimated clusters and outliers.
Comparison of alternative approximate EM methods

Log likelihood (bound) over consecutive iterations, is a run on the synthetic data set. For MCMC using Gibbs sampling, 10,000 samples were used for computing the posterior estimates and the first 2,000 samples were discarded as burn-in.

All algorithms were started from the same parameter values.
Accuracy of detecting ‘genuine’ outliers

- Synthetic data sets generated as before
- Five different error levels defined: The diagonal elements of the variance matrices $S$ will range between $[0,0.01]$, $[0,0.1]$, $[0,1]$, $[0,10]$ and $[0,100]$ respectively.
- Computed the Area Under the ROC curve (AUC) achieved by the inferred outlierness criteria against true outliers, averaged over 10 independent runs.

Further tests on semi-synthetic data, derived from real data with more realistic density structure. It still allows us to evaluate the benefits of having measurement error information, in a controlled manner
  - The ‘lymphography’ (148 data points, 18 dimensions). Initially 4 classes, of which two are too small (6 points in total) and will be considered outliers.
  - We simulate measurement errors by adding hetero-schedastic Gaussian noise with variances ranging in $[0,0.1]$
Accuracy of detecting “genuine” outliers: Results on the synthetic data sets

In sample

Out of sample

![Graphs showing AUC vs Noise level for In sample and Out of sample data sets.](image-url)
Accuracy of detecting “genuine” outliers: Results on semi-synthetic data sets

ROC curves, measured on out-of-sample data, averaged over 10 runs
Application to Astrophysical Data Analysis

- In astrophysical measurements, there is no error-free data
- But estimated measurement errors are available, from knowledge about observing conditions and instrumental limitations.
- We analysed a data set of 10,000 quasars extracted from a well studied survey in astrophysics: the SDSS quasar catalogue
- Features: 5 magnitudes measured with five different optical filters (u,g,i,r,z). From these, to avoid bias with brightness, we construct 4 features, each related to a color, by subtracting r from the others.
- Spectroscopic redshift estimates are available, which we use for validating the interpretability of our results. There is physical reason for high redshift objects to be perceived as outliers in the density of quasars in the color space
Application: Detecting high-redshift quasars from the SDSS quasar catalog

AUC vs. different redshift thresholds. A large fraction of quasars at redshift 2.5 or higher are detected with high probability.

This is not possible with 2D projections as currently used by astronomers.
Conclusions

- We developed a robust mixture model for multivariate data that includes error information.
- We employed a structured variational EM, which in the zero limit of measurement errors reduces to ML estimation of t-mixtures.
- Empirical results show systematic and statistically significant improvements in terms of correct outlier detection rates in high measurement uncertainty conditions.
- An application to detecting high redshift quasars from the SDSS photometric quasar catalogue was demonstrated.
- We are extending these findings to combined statistical and visual analysis of data with outliers and error information.
Extension to constrained mixtures

- Generative topographic mapping for data with outliers and measurement errors

\[ p(t; W, \sigma, \nu) = \frac{1}{K} \sum_{k} \int \int \mathcal{N}(t|w, S) \mathcal{N}(w|W \phi(x_k), \frac{1}{u\sigma}) G(u|\frac{\nu_k}{2}, \frac{\nu_k}{2}) du dw \]

- Despite the continuous latent space being discretised, this model is still intractable

- However, the variational estimation methodology presented for unconstrained mixtures can be adapted for the case of constrained mixtures.

- The non-outliers are assumed to lie on a 2D manifold. The outliers have no structure assumed. Measurement errors distort the overall configuration, but their variances are known.
Some results

• Synthetic data in 10-D: 3 clusters & outliers
• It is beneficial to represent the outlierness in a 3rd dimension.
• Left: visualising a held-out set
• Right: evaluation of KNN classifiers on the low-dimensional coordinates, with (red) and without (green) taking into consideration the measurement error information.
Analysis of SDSS data is underway

• We try to combine statistically sound density-based analysis with visual presentation of the density structure.

A subset of 2,000 quasars from the SDSS
• Left: Similarly to unconstrained mixtures, the constrained mixture is also able to infer outlierness that relates to redshift.

• Right: Outlierness estimates with and without taking measurement errors into account do differ.

The available errors associated with experimental data sets can be exploited for locating ‘interesting’ outliers that are atypical for a reason other than measurement errors.
Part II: On Bayesian classification with Laplace priors
Introduction

• The Laplace density has been widely known and used as a sparsity-inducing prior
• All these works compute MAP estimates
• How about a Bayesian estimate?

Previous attempts
• (Ju, Madigan & Scott, 2002, unpublished)
  – take a Gibbs sampling approach
  – Results are contradictory and inconclusive
• (Park & Casella, 2006, unpublished)
  – Gibbs sampling employed
  – Presents results for the overdetermined regression case only
The model of probit regression

- Training set \( \{(x_1,z_1),\ldots,(x_N,z_N)\} \) where \( x_n \) are \( T \)-dimensional input points; \( z_n \) are labels in \( \{-1,1\} \).
- Classification aims at learning a mapping from inputs to target labels, that is able to predict the label for new inputs.
- Consider a linear regression likelihood model first, with continuous targets \( y=(y_1,\ldots,y_N) \)

\[
y|x,w \sim \mathcal{N}(y|x^Tw,\sigma^2I)
\]

where \( x \) is \( N \times T \) matrix and \( z \) is \( N \times 1 \) vector. and there is a bias term implicit in this notation. Assume \( \sigma = 1 \)
- Now, regard \( y \) as a latent variable & employ the probit link

\[
z_n|y_n \sim P(z_n = 1|y_n) = P(y_n \geq 0) = \Phi(x_n^Tw), \forall n = 1,\ldots,N
\]

where \( \Phi(y) = \int_{-\infty}^{y} \mathcal{N}(u|0,1)du \)
... with Laplace priors

- The prior on $\mathbf{w}$ is chosen to be Laplacian
  \[ \mathbf{w} \sim \frac{\sqrt{\lambda}}{2} e^{-\sqrt{\lambda}|\mathbf{w}|} \]

- The Laplace density is strongly peaked at zero, which should express the prior belief that the feature relevancies are strongly peaked at zero.

- In regression, the use of Laplace priors leads to LASSO

- The Laplace density is a scale-mixture, and this property was exploited in (Figueiredo, 2003) for deriving an iterative E-M algorithm for computing the MAP estimates

\[
\begin{align*}
\mathbf{w}_t | \tau_t & \sim \mathcal{N}(\mathbf{w}_t | 0, \tau_t) \\
\tau_t | \lambda & \sim \mathcal{G}(\tau_t | 1, \lambda/2) = \frac{\lambda}{2} e^{-\lambda \tau_t / 2}
\end{align*}
\]
Bayesian analysis: Insights

- In the 1D case, the convolution of a Gaussian likelihood with a Laplace prior is analytically computable.

  Posterior mean & one posterior std on both sides (vertical axis), against fixed equidistant values of $y_{1,\ldots,N}$ in a range, and having set all 1D inputs $x_{1,\ldots,N}$ to 1.

  Equivalently, we can regard it as the multivariate posterior mean vector of $w$ (vertical axis) against $w^{\text{MaxLikelihood}}$ (horizontal axis).

- Looks quite different from the known sparsifying shrinkage of the MAP estimator...
A variational Bayesian solution

- In the more general multivariate case, the model is not tractable analytically. We develop a variational solution.
- This will enable us to study the practical implications of the posterior induced by the Laplace prior.

- Lower bound the regression likelihood:

\[
\log p(y|x) = \log \int dwd\tau p(y|w, x)p(w|\tau) \prod_t p(\tau_t) \\
\geq \int dwd\tau q(w, \tau) \log \frac{p(y|w, x)p(w|\tau) \prod_t p(\tau_t)}{q(w, \tau)}
\]

where \( q(w, \tau) = q(w) \prod_t q(\tau_t) \) is the variational posterior.

- This decouples is separate terms, and we can compute the variational posteriors.
• The variational posteriors:

\[
q(w) \propto \exp \int d\tau q(\tau) \log \mathcal{N}(y | x^T w, I / \sigma^2) \mathcal{N}(w | 0, \Lambda)
\]

\[
= \mathcal{N}(w | \mu_w, \Sigma_w)
\]

where \( \Lambda = \text{diag}(1/\tau_i) \); \( \langle \Lambda \rangle = \text{diag}(\langle 1/\tau_i \rangle) \) and

\[
\mu_w = \sigma^{-2} \Sigma_w x y = \langle w \rangle
\]

\[
\Sigma_w = \left\{ \langle \Lambda \rangle + \sigma^{-2} x x^T \right\}^{-1} = \langle w w^T \rangle - \langle w \rangle \langle w \rangle^T
\]
• Further,

\[ q(\tau_t) \propto \exp \int dw_t q(w_t) \log \mathcal{N}(w_t|0, \tau_t)Ga(\tau_t|1, \frac{\lambda}{2}) \]

\[ \propto \mathcal{N}(\sqrt{\langle w_t^2 \rangle}|0, \tau_t) \exp(0.5\lambda\tau_t) \]

with the normalization term being:

\[ \int d\tau_t \mathcal{N}(\sqrt{\langle w_t^2 \rangle}|0, \tau_t)Ga(\tau_t|1, \frac{\lambda}{2}) = \frac{\sqrt{\lambda}}{2} \exp \left\{ -\sqrt{\lambda\langle w_t^2 \rangle} \right\} \]

And we can compute the expectation required in the posterior statistics of \( \mathbf{w} \) earlier

\[ \langle 1/\tau_t \rangle = \int d\tau_t \frac{1}{\tau_t} q(\tau_t) = \sqrt{\frac{\lambda}{\langle w_t^2 \rangle}} \]

The difference from the expressions derived in (Figueiredo, 2003) is that \( \langle \mathbf{w}^2 \rangle \) appear in place of \( \langle \mathbf{w} \rangle^2 \)
Finally, for probit classification we also have $q(y)$, which is a product of truncated Gaussians whose expectation is

$$
\langle y \rangle = x^T \langle w \rangle + z \frac{N(x^T \langle w \rangle | 0, 1)}{\Phi(zx^T \langle w \rangle)}
$$

Since $y$ is a hidden variable vector in this case, it needs to be integrated from $q(w)$, which boils down to replacing $y$ by $\langle y \rangle$ in the previous posterior computations.
Inspecting the posterior approximation

- Using the feasible univariate case, we can plot the true (dotted line) vs approximate (continuous line) posteriors. Left: $y=0.5$; Right $t=1$; Both: $\lambda=1$. Note, the mode is at zero on both plots, while the mean is not.
The MAP method: Whence the sparsity?

• It is more convenient to work with the Laplace prior directly here
• MAP maximises the joint probability:

\[ w^* = \arg\max_w \left( -\sum_n \frac{(y_n - x_n^T w)^2}{2\sigma^2} - \sqrt{\lambda}|w| \right) \]

• This is non-differentiable at zero.
Consider the positive case first, $w_t > 0$. Taking derivatives on the positive domain and equating to zero, we obtain:

\[
0 = \frac{\partial \log p(w|y, x)}{\partial w_t} \\
= \sum_n (y_n - \sum_{t' \neq t} w_{t'} x_{t'n} - w_t x_{tn}) x_{tn} / \sigma^2 - \sqrt{\lambda} \\
= \sum_n (y_n - \sum_{t' \neq t} w_{t'} x_{t'n} ) x_{tn} / \sigma^2 - w_t \sum_n x_{tn}^2 / \sigma^2 - \sqrt{\lambda}
\]

\[
w_t = \frac{\sum_n (y_n - \sum_{t' \neq t} w_{t'} x_{t'n} ) x_{tn}}{\sum_n x_{tn}^2} - \sigma^2 \sqrt{\lambda} / \sum_n x_{tn}^2
\]

\[
\underbrace{w_t^{\text{MaxLikelihood}}}_{W_t}
\]

Since we assumed $w_t > 0$, the above equation only admits a solution if

\[
w_t^{ML} > \frac{\sqrt{\lambda} \sigma^2}{\sum_n x_{tn}^2}
\]
Analogously, on the strictly negative domain, \( w_t < 0 \), we get that a solution exists only if \( w_t^{ML} < -\frac{\sqrt{\lambda_2} \sigma^2}{\sum_n x_{2t}^2} \). Thus, in all other cases (i.e. when the ML solution lies between these two thresholds), the solution must be exactly zero.

- In summary, for the unique maximum argument \( w_t^* \), we have:

\[
    w_t^* = \begin{cases} 
    w_t^{ML} - \epsilon_t \text{sign}(w_t^{ML}), & \text{when } |w_t^{ML}| > \epsilon_t \\
    0, & \text{when } |w_t^{ML}| \leq \epsilon_t
    \end{cases}
\]

where \( \epsilon_t = \frac{\sqrt{\lambda_2} \sigma^2}{\sum_n x_{2t}^2} \)

In addition \( \lambda \) may be varied to further control the sparsity level.
Hyperparameter inference

• Irrespective of the estimation / inference method used, $\lambda$ controls the amount of shrinkage globally.

• Cross-validation over a grid of values is computationally demanding

• We place a Gamma hyperprior on this parameter

$$\lambda \sim Ga(\lambda | \alpha, \beta)$$

with $\alpha = \beta = 1$ far enough from zero to avoid problems of vague priors in small sample size conditions. In other words, $\lambda$ can vary cf. an exponential hyper-prior with mean of 1.

• The model and algorithm presented previously is extended to accommodate this hyperprior.
• New variational posterior perm:

\[
q(\lambda) \propto \exp \int d\tau_t q(\tau_t) \sum_t \log \{p(\tau_t | \lambda)p(\lambda | \alpha, \beta)\}
\]

\[
= \mathcal{G}a(\lambda | \alpha + T, \beta + \frac{1}{2} \sum_t \langle \tau_t \rangle)
\]

• Previous occurrences of \( \lambda \) get replaced with the posterior expectation \( \langle \lambda \rangle \) w.r.t. this variational posterior. This is:

\[
\langle \lambda \rangle = \frac{2(\alpha + T)}{2\beta + \sum_t \langle \tau_t \rangle}
\]

• For computing it we also need:

\[
\langle \tau_t \rangle = \int d\tau_t q(\tau_t) = \frac{1 + \sqrt{\langle w_t^2 \rangle} \langle \lambda \rangle}{\langle \lambda \rangle}
\]

\[
= \frac{1}{\langle \lambda \rangle} + \sqrt{\frac{\langle w_t^2 \rangle}{\langle \lambda \rangle}} = \frac{1}{\langle \lambda \rangle} + \langle 1/\tau_t \rangle^{-1}
\]

The algorithm is then to iterate the updating of all required posterior statistics until convergence.
A possible extension: feature-specific hyperparameters

- Similarly to the generalised LASSO, we can have separate hyperpriors for each feature.
- Of course, in that case the prior on \( w \) is no longer Laplacian but the convolution of a Laplace with a Gamma. This behaves similarly to a Student prior if the Gamma is vague (if \( \alpha = \beta = 10^{-6} \)), i.e. the RVM.

\[ \lambda_t \sim Ga(\lambda_t | \alpha, \beta) \]

Then we have

\[ q(\lambda_t) = Ga(\lambda_t | \alpha + 1, \beta + \frac{1}{2} \langle \tau_t \rangle) \]

and so,

\[ \langle \lambda_t \rangle = \frac{2(\alpha + 1)}{2\beta + \langle \tau_t \rangle}; \quad \langle \tau_t \rangle = \frac{1}{\langle \lambda_t \rangle} + \langle 1/\tau_t \rangle^{-1} \]
Related methods

- (Figueiredo, 2003): Jeffrey’s prior on $\tau_t$, which gets rid of hyperparameters.
- The relevance vector machine (RVM) uses independent Student t priors on $w$.
- Difficult to study the relationships between methods on the modelling levels, because the effects of prior definitions and those of posterior approximations interfere.
- We can graphically characterise the joint effect of these factors by visualising the shrinkage functions produced, by simulations (fixing $x=I$ and plotting the estimated posterior statistics of $w$ against the true values of $w$).
• (Variational) posterior mean shrinkage:

solid: VB, Laplace(\(\lambda\))
dash-dot: MAP, Laplace(\(\lambda\))
dash: RVM – Student\(10^{-6}, 10^{-6}\);
ext-Laplace&Ga\(10^{-6}, 10^{-6}\)
dotted: Jeffreys
thin dash: \(w_{est} = w_{true}\) (OLS)
• Var osterior means & one var posterior std on both sides:

Top left: Laplace VB(1)
Top right: RVM indistinguishable from Laplace+Gamma (extension)
Bottom left: Laplace MAP
Bottom right: Gauss+Jeffreys
Predictive distributions

- Given a test point $x^*$, the predictive distribution of its associated target, $y^*$, is the following:
  \[ p(y^*|x^*) = \int dwd\tau s \lambda p(y^*|x^Tw)q(w)q(\tau)q(\lambda) \]
  \[ = \int d\tau d\lambda N(y^*|x^T\mu_w, x^*\Sigma_w x^T + \sigma^2 I)q(\tau)q(\lambda) \]

- It is common to use
  \[ p(y^*|x^*) \approx N(\mu_{y^*}, \nu_{y^*}) \]

where

\[ \mu_{y^*} = x^*\sigma^{-2}\{\langle \Lambda \rangle + \sigma^{-2}x^Tx\}^{-1}x^Ty \]
\[ \nu_{y^*} = x^*(\langle \Lambda \rangle + \sigma^{-2}xx^T)^{-1}x^T + \sigma^2I \]

Further for probit-regression, we have

\[ P(z = 1|x^*) = \int dy^* P(z = 1|y^*)p(y^*|x^*) = \Phi \left( \frac{\mu_{y^*}}{\sqrt{\nu_{y^*}}} \right) \]
Simulation results

• We are interested in the underdetermined case
  – This is when the prior has the largest influence
  – This is the practically relevant case for applications to gene expression classification

• Synthetic data experiments—aimed at finding out which method and which estimation scheme is suited to which data characteristics
  – I.I.d. features
  – Correlated features
  – Evaluating
    • Misclassification rates
    • Brier scores (takes the prediction uncertainly into account)

• Microarray gene expressions
  – Evaluating the Area under the ROC curve (AUC)
- I.I.d. features.
- Comparative evaluation wrt. misclassification rates
- Laplace-VB is better suited in not extremely sparse cases
- Hyperparameter inference doesn’t influence performance
- Gauss+Jeffreys is quasi-random, tends to switch off all weights
- Brier score results from the same set of experiments
- Laplace-VB has an advantage over ‘sparse’ methods in terms of representing the uncertainty
- The hyperparameter inference further improves on this
• Correlated features
• Laplace-VB has advantage in case of correlated features
• As before, hyperparameter inference further improves the uncertainty estimates
Trying to understand these behaviours

Why is Laplace-VB not suited when very few features are relevant?

- We have seen the Bayesian posterior expectations induced by the Laplace priors are not sparse. The variational Bayesian estimates are not sparse either.

Whence the advantage of Laplace-VB with correlated features?

- Laplace-VB doesn’t discard correlated features while ‘sparse’ methods do:
Colon cancer prediction

- 2000 genes; 40 tumour + 12 normal tissues.
- Not clear what fraction of genes is relevant
- It has been widely believed that very few genes are relevant
- Yet, recent results in the literature have found improved results with larger sets of genes
- [We needed to try to understand our methods before trying to understand the data]

- Split the data randomly in 40 training and 12 test samples
- Pre-processing: (adopted from Chu et al, 2005) Each gene (feature) of the normalised (zero mean, unit variance) training set is tested for differential expression with the Wilcoxon Rank Sum test, at significance level $p=0.01$. Genes not found differentially expressed at this stage, are discarded from both training and test sets.
• We perform 100 independent bootstrap repeats (train-test splits) using Laplace-MAP, each time determining the hyperparameter by internal 5-fold cross-validation.

• We do 500 independent bootstrap repeats using Laplace-VB, with hyperparameter inference.

• These are the average and std (in brackets) obtained:

<table>
<thead>
<tr>
<th></th>
<th>PROBIT</th>
<th>La-MAP 100 rep</th>
<th>La((\lambda))-VB 500 rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>false +ves (#)</td>
<td>0.660 (0.074)</td>
<td>0.678 (0.034)</td>
<td></td>
</tr>
<tr>
<td>false -ves (#)</td>
<td>1.410 (0.108)</td>
<td>1.270 (0.042)</td>
<td></td>
</tr>
<tr>
<td>error rate (#)</td>
<td>2.070 (0.122)</td>
<td>1.948 (0.050)</td>
<td></td>
</tr>
<tr>
<td>error rate (%)</td>
<td>17.250 (1.014)</td>
<td>16.233 (0.414)</td>
<td></td>
</tr>
<tr>
<td>AUROC (\times 100)</td>
<td>87.976 (1.059)</td>
<td>88.225 (0.482)</td>
<td></td>
</tr>
</tbody>
</table>
Previous results on the same data are comparable to ours:

- (Chu et al., 2005): best 10-fold CV results on the identically preprocessed data: 15.19 \(\pm\) 13.65\%; 26 genes, using probit Gaussian Process with Gamma priors on lengthscales, hyperparameter estimation by MLII, and a gene ranking scheme involving data resampling to aid stability.

- (Shevade & Kheerty, 2003): 17.7% misclassifications, using sparse logistic regression and data resampling.

- (Qi et al., 2004): best results 1.63 \(\pm\) 0.11 misclassifications; 156.76 \(\pm\) 11.86 selected genes, using predictive-ARD-EP.

- (Li et al., 2002): 2.90 \(\pm\) 0.13% misclassifications; 8.15 \(\pm\) 0.13 genes, using RVM with MLII and some speed-up heuristics.

- (Guyon et al., 2002): 2.94 \(\pm\) 0.14 misclassifications; 4.25 \(\pm\) 0.12 genes, using SVM with recursive feature elimination.
Conclusions

• Contrarily to the ‘myth’ of sparse Bayesian learning with Laplace priors, the sparsity effect is a property of the MAP estimates only. Bayesian estimates have other advantageous properties:
  • Provides better estimates of the prediction uncertainty
  • Able to retain correlated features, which favour generalisation
  • More stable w.r.t. the hyperparameter choice
  • Produces a weighted ranking of features, which may be interpreted.
• Laplace-VB is suitable when feature importance is uneven
• It is not suitable when a large fraction of genes is completely irrelevant.
• We are investigating these priors with generative classifiers on gene expressions...
... and finding interesting results

Leukaemia

Colon

Prostate

Lymphoma

SBRCT

Brain

Breast

Ovarian
Acknowledgements

• Funding from PPARC (relating to Part I) and a Welcome Trust VIP award (relating to Part II) are gratefully acknowledged.

• The material in Part I is based on joint work with Jianyong Sun and Somak Raychaudhury (School of Physics & Astronomy)