Modelling big data fast and on your laptop via symbolic data analysis

Scott A. Sisson, Boris Beranger, Xuhui Fan, Jaslene Lin
Tom Whitaker, Xin Zhang

UNSW Sydney & ACEMS

Bayes on the Beach 2017, Nov 13-15 2017, Gold Coast
Talk Outline

1. Motivation
2. Existing SDA methods
3. New likelihood models for SDA
4. Examples
5. Discussion
Talk Outline

1. Motivation
2. Existing SDA methods
3. New likelihood models for SDA
4. Examples
5. Discussion
Rise of non-standard data forms

Standard statistical methods analyse classical datasets

E.g. \( x_1, \ldots, x_n \) where \( x_i \in \mathcal{X} = \mathbb{R}^p \)

However: Increasingly see non-standard data forms for analysis.

Simple non-standard forms:

- Can arise as result of measurement process
- Blood pressure naturally recorded as \((\text{low}, \text{high})\) interval
- Particulate matter directly recorded as counts within particle diameter ranges i.e. histogram
Example: Discretised data = histogram

Scatterplot with loess line

- E.g. point \((4.0, 0.0)\) actually lies within \([3.95, 4.05) \times [-0.05, 0.05)\)
- Strong discretisation could have undesired inferential impact
Symbolic Data Analysis

- Established by Diday & coauthors in 1990s.

- Basic unit of data is a distribution rather than usual datapoint.
  - interval $(a, b)$
  - $p$-dim hyper-rectangle
  - histogram
  - weighted list etc.
  - can be complicated by “rules”

- Classical data are special case of symbolic data:
  
  E.g. symbolic interval $s = (a, b)$ equivalent to classical data point $x$ if $x = a = b$.
  
  Or histogram $\rightarrow \{x_i\}$ as $\# \text{ bins} \rightarrow \infty$.

So symbolic analyses must reduce to classical methods.
How do symbolic data arise?

- Can arise naturally (measurement error): E.g. blood pressure, particulate histogram, truncation/rounding.

- ‘Big Data’ context:
  - Symbolic data points can summarise a complex & very large dataset in a compact manner.
  - Retaining maximal relevant information in original dataset.
  - Collapse over data not needed in detail for analysis.
  - Summarised data have own internal structure, which must be taken into account in any analysis.

Statistical question:
How to do statistical analysis for this form of data?
Tick time series data

Too much data to analyse all ticks.

Collapse data to e.g. one histogram per day.

Analysis of histograms now tractable. (Though method perhaps unclear.)

In general: Reduction to symbols is question and data dependent.
Example: Analysis of aerosol particulate data

- Aerosol particulate data over boreal forest site (Hyytiälä, Finland)
- Dataset of particle size measurements (histograms) over time
- Can we model evolution of histograms over time?

Fitting regressions to underlying data while only observing symbols.
Estimating number of species from partial information

- **Data:** expert elicited estimates of number of species from 1950–2015
- Mostly intervals: \([a, b]\)
- Some point estimates: \(x\)
- How to reconcile two forms of data for analysis?
- Could potentially treat \([a, b]\) as random intervals in hierarchical model.
- Not sensible to consider \(x\) as classical limit of \([a, b]\)
- Rather \(x\) is ‘partially observed’ interval (see later).

From Lin et al. (2017).
Prediction of crop types from satellite images

From Whitaker et al. (in preparation).

- \(~250\text{K} \) pixels with 6-dimensional predictor variable $x$
- 7 response categories with known ground truth
- Multinomial logistic regression, but computational to fit (hours)
- Can we use SDA to speed things up while maintaining prediction quality?
Finite mixture models for large datasets

▶ Want to fit mixture model

\[ x_i \sim \sum_{j=1}^{J} w_j N_{d}(\mu_j, \Sigma_j) \]

▶ Typically implemented using indicator variables \( z_i \) so that

\[ x_i | z_i \sim N_{d}(\mu_{z_i}, \Sigma_{z_i}) \]

for \( i = 1, \ldots, n \).

▶ What if \( n \) is big?

- Higgs Dataset: \( n = 11m, d = 21 \)
- Very many \( z_i \) to estimate and mix over
- Computation high
- Expect slow mixing

▶ Can we use SDA to speed this up?
Talk Outline

1. Motivation
2. Existing SDA methods
3. New likelihood models for SDA
4. Examples
5. Discussion
How to analyse symbolic data?

- Poorly developed(?) in terms of inferential methods.
- General approaches:
  - **Descriptive statistics** (means, covariances)
    ⇒ Methods based on $1^{st}/2^{nd}$ moments: clustering, PCA etc.
  - **Least-squares optimisation**
    ⇒ Based on minimising distance errors $D(S_1, S_2)$
    ⇒ Can fit e.g. regressions, but no inference (e.g. hyp. tests)
- **Likelihood methods**:
  - Single technique for constructing likelihood functions
    ⇒ Enables model-based inferences. ✓
- Over-prevalence of models for intervals $(a, b)$ & assuming uniformity
  ⇒ Need to move beyond both of these.

**Current SDA research:**
Developing practical model-based (e.g. likelihood-based) procedures for statistical inference using symbolic data for general symbols.
Existing likelihood models for symbols (Le Rademacher & Billard, 2011)

Symbol: \( S = (S^1, \ldots, S^d)^\top \)

E.g. For random intervals \([a_i, b_i], i = 1, \ldots, n:\)

- \( S_i = (a_i, b_i)^\top \)
- \( S_i = (m_i, \log r_i)^\top \)

Then specify a standard (classical data) model for \( S_1, \ldots, S_n. \) E.g.

\[
(m_i, \log r_i)^\top \sim N(\mu, \Sigma)
\]

Simple, and can be very powerful.

Nice application of this in consistent species diversity estimation →
Estimating number of species from partial information

- **Data**: expert elicited estimates of number of species from 1950–2015
  - Mostly intervals: \([a, b]\)
  - Some point estimates: \(x\)
- How to reconcile two forms of data for analysis?
  - Could potentially treat \([a, b]\) as random intervals in hierarchical model.
  - Not sensible to consider \(x\) as classical limit of \([a, b]\)
  - Rather \(x\) is ‘partially observed’ interval (see later).

From Lin et al. (2017).
Estimating number of species from partial information

- Intervals \([a_i, b_i]\) → \(S_i = (m_i, \log r_i)^\top\)

Points \(x\): \(x \rightarrow S_i = (m_i, \log r_i)^\top\) with \(m_i = x_i\) and \(r_i\) unknown

- Estimate missing \(r_i\) (recovering missing interval) by assuming

\[(m_i, \log r_i)^\top \sim N(\mu, \Sigma)\]

and estimating correlation.

- Hierarchical bivariate Gaussian model:

E.g.

\[
\mu_{\text{beetles}} + \mu_{\text{other–insects}} = \mu_{\text{insects}}
\]

“Similarly” for variances

Common correlation \(\rho\)

From Lin et al. (2017).
Estimating number of species from partial information

- Analysis is fusion of different data types
- Consistent diversity estimates across species
- Sharing of information (hierarchy) provides lots of information to improve estimates.

From Lin et al. (2017).
Existing likelihood models for symbols  (Le Rademacher & Billard, 2011)

Model specification issues:

▶ Need to find credible models for general $S$.
  
  • Not always obvious how to do this.
  • Easy to specify models for classical data (e.g. GEV)
  • But how to develop equivalent models for symbols (with internal variation).
  • Can’t just fit to means. How to account for variation? etc.

Inference issues:

▶ Symbols are summaries of classical data, $S = \varphi(X_1, \ldots, X_N)$
  
  • Inference at symbol level only
  • Model can only predict symbols

▶ This is ok, but what if interest is in modelling underlying data?
  
  • Inferential questions at level of classical data
  • Want full distributional predictions of $x$ (not just mean/var)
Existing likelihood models for symbols (Le Rademacher & Billard, 2011)

Symbol issues:

▶ Sometimes symbol assumptions are not realistic:
  
  - Distribution within interval \([a, b]\) often assumed as uniform
  - Extremely unlikely (and affects inference/prediction)

Some symbol parameterisation issues:

▶ Some symbol modelling parameterisations are not stable:
  
  - E.g. \([a, b] = (m, \log r)^T\) is common
  - But model collapses as \(a \rightarrow b\) (i.e. \(\log r \rightarrow \infty\))

▶ So does \(S \rightarrow x\) (classical) produce classical data likelihood?

Question:

How to fit models and make predictions at the level of the classical data, based on only observing symbols?
Talk Outline

1. Motivation
2. Existing SDA methods
3. New likelihood models for SDA
4. Examples
5. Discussion
One possible approach (Beranger, Lin & Sisson, 2017, in preparation)

The general approach:

\[ L(S|\theta, \phi) \propto \int_x g(S|x, \phi)L(x|\theta)dx \]

where

- \( L(x|\theta) \) – standard, classical data likelihood
- \( g(S|x, \phi) \) – probability of obtaining \( S \) given classical data \( x \)
- \( L(S|\theta) \) – new symbolic likelihood for parameters of classical model

**Gist:** Fitting the standard classical model \( L(x|\theta) \), when the data are viewed only through symbols \( S \) as summaries.

As \( S_i \to x_i \), then \( g(S_i|x, \phi) = g(x_i|x) = \delta_{x_i}(x) \) and so

\[ L(S_i|\theta, \phi) \propto \int_x \delta_{x_i}(x)L(x|\theta)dx = L(x_i|\theta) \] (classical likelihood)

Different symbols give different forms for \( g(S|x, \phi) \) (and \( \therefore L(S|\theta, \phi) \)).
How to construct $g(S|x, \phi)$?

Typically we can easily describe the distribution of $X|S$:

- **Intervals:** $x \sim U(a, b)$ where $S = (a, b)^T$

- **Histograms:** $
  \begin{cases} 
  w_i U(b_i, b_{i+1}) & b_i \leq x \leq b_{i+1} \\
  0 & \text{else} 
  \end{cases}$

  for fixed $\{b_i\}$ where $S = (s_1, \ldots, s_B)^T$, $w_i = s_i / \sum_k s_k$.

- **Gaussian:** working on it . . .

although $U(a, b)$ specifications are unrealistic (we avoid this later).

If we specify a prior/marginal on $S$, we then obtain

$$g(S|x, \phi) = g(S|x) = \frac{f(x|S)f(S)}{f(x)}$$

where $f(x) = \int f(x|S)f(S)dx$.

*Cute for Bayesians* as we use a posterior to build a classical likelihood :-)}
Specific cases (1)

Example (1): No specified generative model $L(x|\theta)$

\[
L(S|\theta, \phi) \propto \int_x g(S|x, \phi)L(x|\theta)dx
\]

\[\Rightarrow L(S|\phi) \propto g(S|\phi)\]

That is:

Directly modelling symbol = existing likelihood approach

(Le Rademacher & Billard, 2011) ✓
Specific cases (2): Random intervals

Example (2): Random intervals: \( S = (S_\ell, S_u) \top \)

Assume:

- \( X_1, \ldots, X_n \sim h(x|\omega) \) for some \( h \) (not uniform!) and
- \( S_\ell = X_{(\ell)} \) and \( S_u = X_{(u)} \) are lower/upper order statistics.

Then density of \( X|S \) is easily specified as:

\[
f(x|s_\ell, s_u) = \prod_{k=1}^{\ell-1} h^{(s_\ell)}(x_k|\omega) \prod_{k=1}^{u-\ell-1} h^{(s_\ell, s_u)}(x_{(\ell+k)}|\omega) \prod_{k=1}^{n-u} h^{(s_\ell)}(x_{(u+k)}|\omega) \delta_{s_\ell}(x_{(\ell)}) \delta_{s_u}(x_{(u)})
\]

where

- \( x = (x_1, \ldots, x_n) \top \)
- \( h^{(s_\ell)}(x|\omega) \propto h(x|\omega)I(x < s_\ell), \ h^{(s_u)}(x|\omega) \propto h(x|\omega)I(x < s_u), \) and
  \( h^{(s_\ell, s_u)}(x_{(\ell+k)}|\omega) = h(x|\omega)I(s_\ell < x < s_u). \)
- Delta functions enforce \( x_{(\ell)} = S_\ell \) and \( x_{(u)} = S_u. \)
Specific cases (2): Random intervals

Now, as \( X_1, \ldots, X_n \sim h(x|\omega) \), we also have

\[
f(s_{\ell}, s_u|\omega) = \frac{n!}{(\ell - 1)!(u - \ell - 1)!(n - u)!} H(s_{\ell}|\omega)^{\ell - 1} \times [H(s_u|\omega) - H(s_{\ell}|\omega)]^{u - \ell - 1} [1 - H(s_u|\omega)]^{n - u} h(s_{\ell}|\omega) h(s_u|\omega)
\]

where \( H(x|\omega) = \int h(z|\omega)dz \).

And so we have the joint distribution as

\[
f(x, s_{\ell}, s_u|\omega) = \frac{n!}{(\ell - 1)!(u - \ell - 1)!(n - u)!} \prod_{k=1}^{n} h(x_{(k)}|\omega) \delta_{s_{\ell}}(x_{(\ell)}) \delta_{s_u}(x_{(u)})
\]

and finally

\[
g(s_{\ell}, s_u|x) = \frac{n!}{(\ell - 1)!(u - \ell - 1)!(n - u)!} \delta_{s_{\ell}}(x_{(\ell)}) \delta_{s_u}(x_{(u)}).
\]

Note:

This is independent of the form of \( h(x|\omega) \)!
Specific cases (2): Random intervals

Now if we want to fit the model \( X_1, \ldots, X_n \sim g(x|\theta) \), this gives us

\[
L(s_\ell, s_u|\theta) \propto \int_x g(s_\ell, s_u|x, \phi) \prod_{k=1}^n g(x_k|\theta) dx
\]

\[
\propto \frac{n!}{(\ell - 1)!(u - \ell - 1)!(n - u)!} G(s_\ell|\theta)^{\ell - 1} [G(s_u|\theta) - G(s_\ell|\theta)]^{u - \ell - 1} \\
\times [1 - G(s_u|\theta)]^{n-u} g(s_\ell|\theta)g(s_u|\theta)
\]

where \( G(x|\theta) = \int g(z|\theta) dz \)

\( \Rightarrow \) the (known) joint distribution of \((\ell, u)\)-th order statistics of \( \{X_k\} \).

When \( S_\ell = \min_k X_k \) and \( S_u = \max_k X_k \):

\[
L(s_1, s_n|\theta) \propto n(n - 1) [G(s_n|\theta) - G(s_1|\theta)]^{n-2} g(s_1|\theta)g(s_n|\theta) \quad s_1 < s_2
\]

\( \Rightarrow \) the (known) joint distribution of min/max of \( \{X_k\} \).

Symbolic \( \to \) Classical check:
If \( S_\ell \to S_u = x \) (with \( n = 1 \)) then \( L(s_\ell, s_u|\theta) = g(x|\theta) \).
Specific cases (3): Random histograms

- Underlying data
  \[ X_1, \ldots, X_n \in \mathbb{R}^p \sim h(x|\omega). \]

- Collected into histogram (random counts) with fixed bins as:
  \[
  S = (s_1, \ldots, s_B)^	op \\
  = (\#X_i \in B_1, \ldots, \#X_i \in B_B)^	op 
  \]
  such that \( \sum_b s_b = n \).

The density of \( X|S \) is

\[
f(x|s) = \prod_{b=1}^{B} \prod_{\ell=1}^{s_b} h^{(b)}(x_{b\ell}|\omega)I(x_{b\ell} \in B_b)
\]

where
- \( x_{b\ell} \) is the \( \ell \)-th observation in bin \( B_b \).
- \( h^{(b)}(x|\omega) \propto h(x|\omega)I(x \in B_b) \).
- Enforces \( s_b \) observations in bin \( B_b \).
Specific cases (3): Random histograms

By construction the (prior) distribution of counts $S = (s_1, \ldots, s_B)^T$ is

$$f(S|\omega) = \frac{n!}{s_1! \ldots s_B!} \prod P_h^B(\omega)^{s_b}$$

where

$$P_h^B(\omega) = \int_{B_b} h(x|\omega) dx$$

is the probability that any $x$ will fall in bin $B_b$.

Consequently

$$f(x, S|\omega) = \frac{n!}{s_1! \ldots s_n!} \prod_{i=1}^{n} h(x_i|\omega) \prod_{b=1}^{B} I \left( \sum_{i=1}^{n} I(x_i \in B_b) = s_b \right)$$
Specific cases (3): Random histograms

As a result

\[ g(S|x) = \frac{n!}{s_1! \ldots s_B!} \prod_{b=1}^{B} I \left( \sum_{i=1}^{n} I(x_i \in B_b) = s_b \right). \]

Now if we want to fit the model \( X_1, \ldots, X_n \sim g(x|\theta) \), this gives us

\[
L(S|\theta) \propto \int_x g(S|x) \prod_{k=1}^{n} g(x_k|\theta) dx
\]

\[
\propto \frac{n!}{s_1! \ldots s_n!} \prod_{b=1}^{B} \left[ P^g_b(\theta) \right]^{s_b}
\]

where \( P^g_b(\theta) = \int_{B_b} g(x|\theta) dx \)

\( \Rightarrow \) generalises univariate result of McLachlan & Jones (1988). ✓
Specific cases (3): Random histograms

- Can recover classical likelihood as $B \to \infty$

$$
\lim_{B \to \infty} L(S|\theta) \propto \lim_{B \to \infty} \frac{n!}{s_1! \ldots s_B!} \prod_{b=1}^{B} \left[ \int_{B_b} g(z|\theta) dz \right]^{s_b}
$$

$$
= L(X_1, \ldots, X_n|\theta)
$$

So recover classical analysis as we approach classical data. ✓

- **Consistency:** Can show that with a sufficient number of histogram bins can perform analysis arbitrarily close to analysis with full dataset.

- **Some approximation of** $L(S|\theta)$ **to** $L(x|\theta)$ **depending on level of discretisation.** Work needed to quantify this.

- More complicated if data are not iid but exchangeable

Zhang & Sisson (2017), Zhang & Sisson (in preparation)
Talk Outline

1. Motivation
2. Existing SDA methods
3. New likelihood models for SDA
4. Examples
5. Discussion
Fitting a GEV

Suppose \( x_1, \ldots, x_n \sim \text{GEV}(\mu, \sigma, \xi) \) for large \( n \).
Create histogram of counts \( s = (s_1, \ldots, s_B) \).
Symbolic log-likelihood function is then

\[
\ell(s|\mu, \sigma, \xi) \propto \sum_{b=1}^{B} s_b \log \left[ F(a_{b+1}|\mu, \sigma, \xi) - F(a_b|\mu, \sigma, \xi) \right]
\]

Note that \( \ell(s|\mu, \sigma, \xi) \) tends to standard likelihood as \( \# \) bins gets large (so 1 or 0 observations per bin)

Computation:

- Optimisation of \( \ell \) (v. quick)
- Creation of histogram \( s \) (slower)

\[\text{good fits with moderate bin numbers}\]
Fitting a GEV

Mean MSE $\times 10^{-3}$ (1000 reps)

<table>
<thead>
<tr>
<th>B</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$\xi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2.977</td>
<td>7.675</td>
<td>4.091</td>
</tr>
<tr>
<td>10</td>
<td>1.385</td>
<td>1.030</td>
<td>0.916</td>
</tr>
<tr>
<td>20</td>
<td>1.278</td>
<td>0.762</td>
<td>0.682</td>
</tr>
<tr>
<td>1000</td>
<td>1.277</td>
<td>0.809</td>
<td>0.662</td>
</tr>
<tr>
<td>Standard</td>
<td>1.268</td>
<td>0.725</td>
<td>0.547</td>
</tr>
</tbody>
</table>

- Use R’s `hist` command to construct histograms, $n = 1,000$
- Use `fgev` command in `evd` package for standard approach
- Accuracy increases with more bins
- Accuracy close to using full dataset with only 20 bins
  (No real advantage to 1000 bins over 20)
## Fitting a GEV

<table>
<thead>
<tr>
<th>Time in seconds</th>
<th>100</th>
<th>1K</th>
<th>10K</th>
<th>100K</th>
<th>1M</th>
<th>10M</th>
<th>100M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>0.018</td>
<td>0.047</td>
<td>0.431</td>
<td>2.860</td>
<td>(*)&amp;</td>
<td>(*)&amp;</td>
<td>(*)&amp;</td>
</tr>
<tr>
<td>Symbolic (total)</td>
<td>0.060</td>
<td>0.062</td>
<td>0.062</td>
<td>0.107</td>
<td>0.247</td>
<td>2.217</td>
<td>42.994</td>
</tr>
<tr>
<td>Symbolic (hist)</td>
<td>0.055</td>
<td>0.057</td>
<td>0.059</td>
<td>0.104</td>
<td>0.243</td>
<td>2.209</td>
<td>42.943</td>
</tr>
<tr>
<td>Symbolic (mle)</td>
<td>0.005</td>
<td>0.005</td>
<td>0.004</td>
<td>0.003</td>
<td>0.004</td>
<td>0.007</td>
<td>0.051</td>
</tr>
</tbody>
</table>

- Standard initially faster than symbolic for small datasets $\sim 1K$
- Symbolic scales much better $> 1K$
- $\ast = f\text{gev}$ crashed on my laptop!
- However, most time for symbolic on histogram construction
- Actual symbolic optimisation super fast (obviously)
- Possible laptop caching problems around 100M
- Faster ways to construct histogram counts than \texttt{hist} for really large datasets (e.g. map-reduce using \texttt{DeltaRho} in R)
Example: Analysis of aerosol particulate data (Zhang & Sisson, in prep.)

Figure 2. An illustration of a new particle formation event at a Boreal Forest site located in Southern Finland. (a) The temporal variation of the particle number size distribution and (b) selected particle number size distributions showing the different stages of the newly formed particle mode from its early stage. Note that this new particle formation occurred on a regional scale over the southern part of Finland. This figure is available in colour online at wileyonlinelibrary.com/journal/env

3.1. Mixture model at a single time point
The density of data ($y$) given by a finite mixture model can be represented by:

$$p(y|\theta) = \sum_{j=1}^{k} \lambda_j f(y|\theta_j)$$

where $k$ is the number of components in the mixture, $\lambda_j$ represents the probability of membership of the $j$th component ($\sum_{j=1}^{k} \lambda_j = 1$), and $f(y|\theta_j)$ is the density function of component $j$ which has parameters $\theta_j$.

As component membership of the data is unknown, a computationally convenient method of estimation for mixture models is to use a hidden allocation process and introduce a latent indicator variable $z_{ij}$, which is used along the lines of a missing variable approach to allocate observations $y_i$ to each component.

In this paper, we transform the aerosol particle size distribution data using a natural logarithm prior to fitting the mixture model (Whitby and McMurry, 1997). In this case, the data ($y_i$) are the natural log of particle diameters (nm), which are assumed to be normally distributed and the parameters ($\theta_j$) to be estimated for each component are therefore the mean ($\mu_j$) and variance ($\sigma_j^2$), along with the component weight ($\lambda_j$). The number of normal components, $k$, is also assumed to be unknown. Priors were:

$$p(\mu_j) \sim N(\xi, \kappa^{-1})$$
$$p(\sigma_j^2) \sim \text{Gamma}(\delta, \beta)$$

Aerosol particulate data over boreal forest site (Hyytiälä, Finland)
Dataset of particle size measurements (histograms) over time
Can we model evolution of histograms over time?
Example: Analysis of aerosol particulate data (Zhang & Sisson, in prep.)

Classical statistical model:

For time (histogram) \( t \),

\[
y_i(t) = \sum_{k=1}^{K} \lambda_k(t) \phi(y_i(t) | \mu_k, \sigma_k)
\]

for \( i = 1, \ldots, n_t \).

- i.e. Data underlying histogram = mixture of normals
- \( K \) constant components over time.
- Weights \( \lambda_k(t) \) vary over time.
- Only histograms observed.
- Fitted model assumes exchangeable \( y_i(t) \) for each \( t \), not iid
Analysis: 3 component mixture  (Zhang & Sisson, in prep.)

- Two components clearly interchanging between 4am–4pm
- One persistent background component
Analysis: 4 component mixture  (Zhang & Sisson, in prep.)

- More involved interchange between components
- One persistent background component
Analysis: 6 component mixture  (Zhang & Sisson, in prep.)

Deconstruction of histogram change over time
Can compute Monte Carlo estimates of likelihood ⇒ BIC estimates:

- 3 components, BIC ≈ 5,329,654
- 4 components, BIC ≈ 5,267,061
- 5 components, BIC ≈ 5,257,971 (same as previous ad-hoc analysis)
- 6 components, BIC ≈ 5,259,559
Prediction of crop types from satellite images

From Whitaker et al. (in preparation).

- ~250K pixels with 6-dimensional predictor variable $x$
- 7 response categories with known ground truth
- Multinomial logistic regression, but computational to fit (hours)
- Can we use SDA to speed things up while maintaining prediction quality?
Prediction of crop types from satellite images

Have \( x_1, \ldots, x_n \sim g(x|\theta) \) for large \( n \) where \( x_i = (x^1_i, \ldots, x^p_i)^\top \in \mathbb{R}^p \).

Create multivariate histogram of counts \( s = (s_1, \ldots, s_B) \) over \( \mathbb{R}^p \).

\[
\log L(s|\theta) \propto \sum_{b=1}^{B} s_b \log \left[ \int_{B_b} g(z|\theta)dz \right]
\]

Comments:

- Multivariate histograms become inefficient as \( p \) gets large
- Constructing histograms more difficult for large \( n, p \) (map-reduce)
- Univariate bin \( B_b = [a_1, a_2] \), \( \int_{B_b} g(z|\theta)dz = G(a_2|\theta) - G(a_1|\theta) \)
  
  For bivariate bin \( B_b = [a_1, b_1] \times [a_2, b_2] \).

\[
\int_{B_b} g(z|\theta)dz = G(b_1, b_2|\theta) - G(a_1, b_2|\theta) - G(b_1, a_2|\theta) + G(a_1, a_2|\theta)
\]

- \( 2^p \) terms in integral – exact calc only practical for moderate \( p \).
Prediction of crop types from satellite images

For the current analysis:

- Closed-form $G$ only available for 0/1 response and single predictor $x$
- We have 7 response and 6-dimensional predictor
- Use quadrature methods for integration, and ideas inspired by composite likelihoods (CL) to avoid 6-dimensional histogram

<table>
<thead>
<tr>
<th>Regression</th>
<th>Comp. Cost</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regress $y$ on 6-d histogram</td>
<td>extreme</td>
<td>no</td>
</tr>
<tr>
<td>Regress $y$ on CL of all 5-d histograms</td>
<td>high</td>
<td>some</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Regress $y$ on CL of all 2-d histograms</td>
<td>low</td>
<td>yes</td>
</tr>
<tr>
<td>Regress $y$ on CL of all 1-d histograms</td>
<td>trivial</td>
<td>yes</td>
</tr>
</tbody>
</table>

- Not true CL setup - parameter estimates are biased in known way
- Bias known for single model - working the CL maths now
- Real Q: Does the bias affect predictions?
Some predictors clearly identify crops even in 1-D (e.g. Bare soil)
Prediction of crop types from satellite images

- Predictive performance excellent for any CL model with $> 3$ bins
- SDA $\rightarrow$ CL speedup = excellent
- 1-d hist (3 bins) $\approx 2.47$ mins versus full classical $\approx 135$ mins
- This for 100K pixels (simulated data).
  SDA computation = static for larger $n$. Classical $\uparrow$ fast.

From Whitaker et al. (in preparation).
Finite mixture models for large datasets

- Want to fit mixture model

\[ x_i \sim \sum_{j=1}^{J} w_j N_d(\mu_j, \Sigma_j) \]

- Typically implemented using indicator variables \( z_i \) so that

\[ x_i | z_i \sim N_d(\mu_{z_i}, \Sigma_{z_i}) \]

for \( i = 1, \ldots, n \).

- What if \( n \) is big?
  - Higgs Dataset: \( n = 11m, d = 21 \)
  - Very many \( z_i \) to estimate and mix over
  - Computation high
  - Expect slow mixing

- Can we use SDA to speed this up?

From X. Fan & Sisson (in preparation).
Finite mixture models for large datasets

If we use histograms:

\[
\log L(s|\theta) \propto \sum_{b=1}^{B} s_b \log \left[ \int_{B_b} g(z|\theta) \, dz \right]
\]

- For all \( x_i \) in bin \( B_b \), we have many \( z_i \)

\[
Z_b = \{ z_i : x_i \in B_b \} \quad |Z_b| = \text{big.}
\]

- Only \( J \) possible values \( z_i \in Z_b \) can take, so collapse these to

\[
(\kappa_1, \ldots, \kappa_J) \sim \text{Multinomial}(|Z_b|; w_1, \ldots, w_J).
\]

- “Distribution on indicators in each bin”

- Suddenly we only have \( J \times B \ll n \) indicators to track

- Can update \((\kappa_1, \ldots, \kappa_J)\) for each bin in one step

- Down side is no longer have Gibbs Sampler – so M.-H. etc
Finite mixture models for large datasets

- $n = 100,000$, 10 components, 100 bins
- Appears to perform well
- Also shown – mixture fitting for random intervals
Finite mixture models for large datasets

- $n = 100,000$, 5 components, 100 sampler iterations.
- $21 \times 21$, $61 \times 61$ and $151 \times 151$ bins
- Slightly blurry for lowest resolution
 Finite mixture models for large datasets

- SDA computation overheads are constant in $n$
- Classical data methods are exponential
- Seems quite promising.
Talk Outline

1. Motivation
2. Existing SDA methods
3. New likelihood models for SDA
4. Examples
5. Discussion
Summary

New approach to likelihoods in SDA:
- Based on fitting underlying (classical) model
- Views latent (classical) data through symbols
- Recovers known existing models for symbols but is more general
- Recovers classical model as $S \rightarrow x$
- Works for general symbols

Currently working on:
- Characterise impact of using symbols on accuracy
  - Trade-off of accuracy vs computation
- Stochastic integration methods when exact integral unavailable
- Design of symbols for best performance
  - Suspect quantiles is better for random intervals than min/max
  - Histogram setting →
How to design symbolic data?

(a) Regular discretisation  
(b) Quantile discretisation  
(c) Tails focused discretisation

How to design symbols to most efficiently represent dataset without (much) loss of critical information?

E. g. Linear regression with 10 million datapoints.
THANK YOU

Some Manuscripts:

https://arxiv.org/abs/1608.00107

https://arxiv.org/abs/1711.03202

▶ Other papers in preparation and to appear on arXiv ‘real soon now’.