InSilicoVA

Samuel Clark

In collaboration with the WHO VA Reference Group
Columbus, Ohio
November, 2018

Outline

Background

InSilicoVA builds on NBC and InterVA to address:

- Uncertainty – do away with 'indeterminate' cause and replace with '(very) uncertain' cause and corresponding uncertainty in related CSMFs

Background

InSilicoVA builds on NBC and InterVA to address:

- Uncertainty – do away with 'indeterminate' cause and replace with '(very) uncertain' cause and corresponding uncertainty in related CSMFs
- Use information on absence of symptoms

Background

InSilicoVA builds on NBC and InterVA to address:

- Uncertainty – do away with 'indeterminate' cause and replace with '(very) uncertain' cause and corresponding uncertainty in related CSMFs
- Use information on absence of symptoms
- Use a variety of SCI from different of sources

Background

InSilicoVA builds on NBC and InterVA to address:

- Uncertainty – do away with 'indeterminate' cause and replace with '(very) uncertain' cause and corresponding uncertainty in related CSMFs

- Use information on absence of symptoms

- Use a variety of SCI from different of sources
  - SCI from physicians (like InterVA)

Background

InSilicoVA builds on NBC and InterVA to address:

- Uncertainty – do away with 'indeterminate' cause and replace with '(very) uncertain' cause and corresponding uncertainty in related CSMFs
- Use information on absence of symptoms
- Use a variety of SCI from different of sources
  - SCI from physicians (like InterVA)
  - SCI in the form of deaths with known cause

Background

InSilicoVA builds on NBC and InterVA to address:

- Uncertainty – do away with 'indeterminate' cause and replace with '(very) uncertain' cause and corresponding uncertainty in related CSMFs

- Use information on absence of symptoms

- Use a variety of SCI from different of sources
    - SCI from physicians (like InterVA)
    - SCI in the form of deaths with known cause
    - De-biased physician-coded VA causes

Background

InSilicoVA builds on NBC and InterVA to address:

- Uncertainty – do away with 'indeterminate' cause and replace with '(very) uncertain' cause and corresponding uncertainty in related CSMFs
- Use information on absence of symptoms
- Use a variety of SCI from different of sources
    - SCI from physicians (like InterVA)
    - SCI in the form of deaths with known cause
    - De-biased physician-coded VA causes

Background

InSilicoVA builds on NBC and InterVA to address:

- Uncertainty – do away with 'indeterminate' cause and replace with '(very) uncertain' cause and corresponding uncertainty in related CSMFs
- Use information on absence of symptoms
- Use a variety of SCI from different of sources
    - SCI from physicians (like InterVA)
    - SCI in the form of deaths with known cause
    - De-biased physician-coded VA causes

InSilicoVA is essentially naive Bayes in a statistical framework (and a bunch of other nice features)

Background

InSilicoVA builds on NBC and InterVA to address:

- Uncertainty – do away with 'indeterminate' cause and replace with '(very) uncertain' cause and corresponding uncertainty in related CSMFs
- Use information on absence of symptoms
- Use a variety of SCI from different of sources
    - SCI from physicians (like InterVA)
    - SCI in the form of deaths with known cause
    - De-biased physician-coded VA causes

InSilicoVA is essentially naive Bayes in a statistical framework (and a bunch of other nice features)

Developed by our team [1, 2]

InSilicoVA Setup

A statistical-computational algorithm that

1. At the individual level identifies a distribution of probabilities associated with each cause

InSilicoVA Setup

A statistical-computational algorithm that

1. At the individual level identifies a distribution of probabilities associated with each cause
2. At the population level identifies a distribution of counts of deaths for each cause

InSilicoVA Setup

A statistical-computational algorithm that

1. At the individual level identifies a distribution of probabilities associated with each cause

2. At the population level identifies a distribution of counts of deaths for each cause

3. Links the two so that they are consistent with each other

InSilicoVA Setup

A statistical-computational algorithm that

1. At the individual level identifies a distribution of probabilities associated with each cause

2. At the population level identifies a distribution of counts of deaths for each cause

3. Links the two so that they are consistent with each other

4. Through these distributions quantifies uncertainty at both levels, again so that each is consistent with the other

InSilicoVA Setup

A statistical-computational algorithm that

1. At the individual level identifies a distribution of probabilities associated with each cause

2. At the population level identifies a distribution of counts of deaths for each cause

3. Links the two so that they are consistent with each other

4. Through these distributions quantifies uncertainty at both levels, again so that each is consistent with the other

InSilicoVA Setup

A statistical-computational algorithm that

1. At the individual level identifies a distribution of probabilities associated with each cause

2. At the population level identifies a distribution of counts of deaths for each cause

3. Links the two so that they are consistent with each other

4. Through these distributions quantifies uncertainty at both levels, again so that each is consistent with the other

Uncertainty: if we are not sure about a cause at the individual level then we cannot be sure about that cause at the population level ...

InSilicoVA algorithm Intuition

Start with the population

- Imagine a vector of 3 CSMFs consisting of decimal numbers that add to 1.0 - e.g. $[0.20, 0.50, 0.30]$

InSilicoVA algorithm Intuition

Start with the population

- Imagine a vector of 3 CSMFs consisting of decimal numbers that add to 1.0 - e.g. $[0.20, 0.50, 0.30]$

- Now imagine a similar but slightly different CSMF vector ... $[0.15, 0.55, 0.30]$

InSilicoVA algorithm Intuition

Start with the population

- Imagine a vector of 3 CSMFs consisting of decimal numbers that add to 1.0 - e.g. $[0.20, 0.50, 0.30]$

- Now imagine a similar but slightly different CSMF vector ... $[0.15, 0.55, 0.30]$

- And now imagine many (e.g. 1,000) slightly different but similar CSMF vectors

InSilicoVA algorithm Intuition

Start with the population

- Imagine a vector of 3 CSMFs consisting of decimal numbers that add to 1.0 - e.g. $[0.20, 0.50, 0.30]$
- Now imagine a similar but slightly different CSMF vector ... $[0.15, 0.55, 0.30]$
- And now imagine many (e.g. 1,000) slightly different but similar CSMF vectors
- These form a distribution of the CSMFs with variation (uncertainty)

InSilicoVA algorithm Intuition

Turning to the individual deaths

- Assume that you have a rough guess of the CSMFs for the population that includes the deaths

InSilicoVA algorithm Intuition

Turning to the individual deaths

- Assume that you have a rough guess of the CSMFs for the population that includes the deaths

- Also assume you have a rough idea of how signs/symptoms (i.e. VA information) are related to causes of death

InSilicoVA algorithm Intuition

Turning to the individual deaths

- Assume that you have a rough guess of the CSMFs for the population that includes the deaths

- Also assume you have a rough idea of how signs/symptoms (i.e. VA information) are related to causes of death

- For a given death with several signs/symptoms pointing to HIV, use what you know (above) and imagine that it is most reasonable to assign this death to HIV

InSilicoVA algorithm Intuition

Turning to the individual deaths

- Assume that you have a rough guess of the CSMFs for the population that includes the deaths

- Also assume you have a rough idea of how signs/symptoms (i.e. VA information) are related to causes of death

- For a given death with several signs/symptoms pointing to HIV, use what you know (above) and imagine that it is most reasonable to assign this death to HIV

- Now introduce a small perturbation in your assumptions and make a new cause assignment for this death; this time it is assigned to TB, closely related and hard to distinguish from HIV

InSilicoVA algorithm Intuition

Turning to the individual deaths

- Assume that you have a rough guess of the CSMFs for the population that includes the deaths

- Also assume you have a rough idea of how signs/symptoms (i.e. VA information) are related to causes of death

- For a given death with several signs/symptoms pointing to HIV, use what you know (above) and imagine that it is most reasonable to assign this death to HIV

- Now introduce a small perturbation in your assumptions and make a new cause assignment for this death; this time it is assigned to TB, closely related and hard to distinguish from HIV

- Now imagine doing this many times (e.g. 1,000) to produce many different but hopefully 'close' cause assignments

InSilicoVA algorithm Intuition

Bring the population and the individual deaths together

- We need the population-level CSMFs and the individual cause assignments to be consistent with one anther ...

InSilicoVA algorithm Intuition

Bring the population and the individual deaths together

- We need the population-level CSMFs and the individual cause assignments to be consistent with one anther ...

- And, we need the uncertainty inherent in cause assignment to be reflected in error bounds on the CSMFs

InSilicoVA algorithm Intuition

Bring the population and the individual deaths together

- We need the population-level CSMFs and the individual cause assignments to be consistent with one anther ...

- And, we need the uncertainty inherent in cause assignment to be reflected in error bounds on the CSMFs

- Imagine a joint distribution of CSMF vectors and individual cause assignments – potentially many dimensions, one for the CSMF vector and one for each individual death, or alternatively for a vector of cause assignments with one element per death

InSilicoVA algorithm Intuition

Bring the population and the individual deaths together

- We need the population-level CSMFs and the individual cause assignments to be consistent with one anther ...

- And, we need the uncertainty inherent in cause assignment to be reflected in error bounds on the CSMFs

- Imagine a joint distribution of CSMF vectors and individual cause assignments – potentially many dimensions, one for the CSMF vector and one for each individual death, or alternatively for a vector of cause assignments with one element per death

- This distribution describes how likely it is that a given CSMF vector is associated with a pattern of cause assignments across all individuals in the population

InSilicoVA algorithm Intuition

Bring the population and the individual deaths together

- We need the population-level CSMFs and the individual cause assignments to be consistent with one anther ...

- And, we need the uncertainty inherent in cause assignment to be reflected in error bounds on the CSMFs

- Imagine a joint distribution of CSMF vectors and individual cause assignments – potentially many dimensions, one for the CSMF vector and one for each individual death, or alternatively for a vector of cause assignments with one element per death

- This distribution describes how likely it is that a given CSMF vector is associated with a pattern of cause assignments across all individuals in the population

- You can easily imagine that only a few relatively similar CSMF vectors are compatible with a given pattern of cause assignments

InSilicoVA algorithm Intuition

How do we do this?

- *Problem* – we cannot observe either of the components of this distribution: CSMFs or cause assignments

How do we do this?

- *Problem* – we cannot observe either of the components of this distribution: CSMFs or cause assignments
- *Solution*

InSilicoVA algorithm Intuition

How do we do this?

- *Problem* – we cannot observe either of the components of this distribution: CSMFs or cause assignments
- *Solution*
  - Model both the CSMFs and the cause assignments so that they are consistent with one another – use VA signs/symptoms and knowledge of the relationship between those and causes of death

InSilicoVA algorithm Intuition

How do we do this?

- *Problem* – we cannot observe either of the components of this distribution: CSMFs or cause assignments
- *Solution*
  - Model both the CSMFs and the cause assignments so that they are consistent with one another – use VA signs/symptoms and knowledge of the relationship between those and causes of death
  - Use this model to characterize the joint distribution of CSMFs and cause assignments

InSilicoVA algorithm Intuition

How do we do this?

- *Problem* – we cannot observe either of the components of this distribution: CSMFs or cause assignments
- *Solution*
  - Model both the CSMFs and the cause assignments so that they are consistent with one another – use VA signs/symptoms and knowledge of the relationship between those and causes of death
  - Use this model to characterize the joint distribution of CSMFs and cause assignments
  - This distribution cannot be defined analytically, but we can approximate it by taking a sample

InSilicoVA algorithm Intuition

How do we do this?

- *Problem* – we cannot observe either of the components of this distribution: CSMFs or cause assignments
- *Solution*
    - Model both the CSMFs and the cause assignments so that they are consistent with one another – use VA signs/symptoms and knowledge of the relationship between those and causes of death
    - Use this model to characterize the joint distribution of CSMFs and cause assignments
    - This distribution cannot be defined analytically, but we can approximate it by taking a sample
    - Marginal distributions from this sample approximate the distributions of CSMFs and cause assignment for individual deaths

Symptom-Cause Information in InSilicoVA

SCI appear in InSilicoVA in exactly the same way as it does in naive Bayes – in the form of $\Pr(s|c)$ in the full naive Bayes relationship,

$$\Pr(c|\mathbf{s}) = \frac{\Pr(c) \prod_s \Pr(s|c)^s (1 - \Pr(s|c))^{1-s}}{\sum_c \Pr(c) \prod_s \Pr(s|c)^s (1 - \Pr(s|c))^{1-s}}$$

The $\Pr(s|c)$ can come from either physicians or a 'gold standard' dataset, *including de-biased physian-coded VAs*

Symptom-Cause Information in InSilicoVA

SCI appear in InSilicoVA in exactly the same way as it does in naive Bayes – in the form of $\Pr(s|c)$ in the full naive Bayes relationship,

$$\Pr(c|\mathbf{s}) = \frac{\Pr(c) \prod_s \Pr(s|c)^s (1 - \Pr(s|c))^{1-s}}{\sum_c \Pr(c) \prod_s \Pr(s|c)^s (1 - \Pr(s|c))^{1-s}}$$

The $\Pr(s|c)$ can come from either physicians or a 'gold standard' dataset, *including de-biased physian-coded VAs*

InSilicoVA can also use other kinds of SCI such as deaths with VA symptoms and a cause assigned through another mechanism

InSilicoVA summary

- Provides estimates of **both** individual-level probabilities of dying from each cause and population level CSMFs

InSilicoVA summary

- Provides estimates of **both** individual-level probabilities of dying from each cause and population level CSMFs
- Both come from the same joint distribution so they are by definition consistent with one another

InSilicoVA summary

- Provides estimates of **both** individual-level probabilities of dying from each cause and population level CSMFs

- Both come from the same joint distribution so they are by definition consistent with one another

- Both are reported as distributions so there is a metric of uncertainty for both

InSilicoVA summary

- Provides estimates of **both** individual-level probabilities of dying from each cause and population level CSMFs

- Both come from the same joint distribution so they are by definition consistent with one another

- Both are reported as distributions so there is a metric of uncertainty for both

- It builds on InterVA and utilizes the same data and expert knowledge, i.e. **if you can run InterVA4 now, then you can also run InSilicoVA**

InSilicoVA summary

- Provides estimates of **both** individual-level probabilities of dying from each cause and population level CSMFs

- Both come from the same joint distribution so they are by definition consistent with one another

- Both are reported as distributions so there is a metric of uncertainty for both

- It builds on InterVA and utilizes the same data and expert knowledge, i.e. **if you can run InterVA4 now, then you can also run InSilicoVA**

- There is an open-source, publicly available R package to run it
  https://cran.r-project.org/package=InSilicoVA

InSilicoVA summary

- Provides estimates of **both** individual-level probabilities of dying from each cause and population level CSMFs

- Both come from the same joint distribution so they are by definition consistent with one another

- Both are reported as distributions so there is a metric of uncertainty for both

- It builds on InterVA and utilizes the same data and expert knowledge, i.e. **if you can run InterVA4 now, then you can also run InSilicoVA**

- There is an open-source, publicly available R package to run it
  https://cran.r-project.org/package=InSilicoVA

- Paper published in a respected statistics journal (JASA) that describes it fully and fairly compares its performance to other algorithms [2]

[1] Samuel J Clark, Tyler McCormick, Zehang Li, and Jon Wakefield. Insilicova: A method to automate cause of death assignment. Working Paper 133, Center for Statistics and the Social Sciences (CSSS), University of Washington, https://www.csss.washington.edu/Papers/wp133.pdf, 2013.

[2] Tyler H McCormick, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark. Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049, 2016.

Extra Slides

InSilicoVA algorithm – Sampling Intuition

Sample the joint distribution of CSMFs and individual cause assignments:

1. Select initial CSMFs - just a place to start

2. **Using naive Bayes relationship and SCI**, calculate the probabilities of the **symptoms recorded for each death**, and use those along with the current CSMFs (from previous step) to draw a new cause for each death from multinomial distribution (can mix in some deaths with known cause)

3. Count the current deaths by cause, and use those counts to draw a new set of CSMFs from Dirichlet distribution

4. Repeat many times ...

InSilicoVA algorithm – Sampling Intuition

Sample the joint distribution of CSMFs and individual cause assignments:

1. Select initial CSMFs - just a place to start

2. **Using naive Bayes relationship and SCI**, calculate the probabilities of the **symptoms recorded for each death**, and use those along with the current CSMFs (from previous step) to draw a new cause for each death from multinomial distribution (can mix in some deaths with known cause)

3. Count the current deaths by cause, and use those counts to draw a new set of CSMFs from Dirichlet distribution

4. Repeat many times ...

With enough repetitions, the sample is a good representation of the joint distribution

Step 2 is where observed symptoms and SCI affect the logic of InSilicoVA

InSilicoVA algorithm - equations and computation 1

Notation

- Deaths: $y_j \in \{0,1\}$ $j \in \{1,\ldots,J\}$
- Sign/symptoms: $s_k \in \{0,1\}$ $k \in \{1,\ldots,K\}$
- Causes: c_n $n \in \{1,\ldots,N\}$
- For individual j, probability of cause n:
  ℓ_{jn} $j \in \{1,\ldots,J\}, n \in \{1,\ldots,N\}$
- Cause-specific death count (CSDC): m_n $n \in \{1,\ldots,N\}$
- Cause-specific mortality fraction (CSMF): f_n $n \in \{1,\ldots,N\}$

InSilicoVA algorithm - equations and computation 2

Data

- For each death y_j, VA interview produces a binary-valued vector of signs/symptoms

$$\{s_{j1}, s_{j2}, \ldots s_{jk}\}$$

InSilicoVA algorithm - equations and computation 2

Data

- For each death y_j, VA interview produces a binary-valued vector of signs/symptoms

$$\{s_{j1}, s_{j2}, \ldots s_{jk}\}$$

- A fixed *probbase*, a $K \times N$ matrix of conditional probabilities derived either from physicians' opinions about how likely a given sign/symptom is for a death resulting from a given cause, **or** from a 'gold standard' dataset containing VA symptoms and causes of death assigned through an independent mechanism (e.g. de-biased physician coding)

$$\begin{bmatrix} \Pr(s_1|c_1) & \Pr(s_1|c_2) & \cdots & \Pr(s_1|c_N) \\ \Pr(s_2|c_1) & \Pr(s_2|c_2) & \cdots & \Pr(s_2|c_N) \\ \vdots & \vdots & \ddots & \vdots \\ \Pr(s_K|c_1) & \Pr(s_K|c_2) & \cdots & \Pr(s_K|c_N) \end{bmatrix}$$

Model

- We are interested in the joint distribution $(\vec{F}, \vec{Y})$

Model

- We are interested in the joint distribution $(\vec{F}, \vec{Y})$
- Model $\vec{F}$ conditional on $\vec{Y}$

$$\vec{F} | \vec{Y} \sim \text{Dirichlet}(\vec{M}) \tag{1}$$

where

$$m_n = \sum_{j=1}^{J} [y_j = c_n], \text{ using Iverson's bracket notation:} [\alpha] = \begin{cases} 1, \alpha \text{ true} \\ 0, \alpha \text{ false} \end{cases}$$

which ties the CSMFs $(\vec{F})$ to the individual cause assignments

InSilicoVA algorithm - equations and computation 3

Model

- We are interested in the joint distribution $(\vec{F}, \vec{Y})$
- Model $\vec{F}$ conditional on $\vec{Y}$

$$\vec{F}|\vec{Y} \sim \text{Dirichlet}(\vec{M}) \tag{1}$$

where

$$m_n = \sum_{j=1}^{J} [y_j = c_n], \text{ using Iverson's bracket notation:} [\alpha] = \begin{cases} 1, \alpha \text{ true} \\ 0, \alpha \text{ false} \end{cases}$$

which ties the CSMFs $(\vec{F})$ to the individual cause assignments

- Model y_j conditional on $\vec{F}$ and $\vec{S}_j$:

$$y_j|(\vec{F}, \vec{S}_j) \sim \text{Multinomial}(\vec{L}_j) \tag{2}$$

Using Bayes' Rule the components of $\vec{L}_j$ in equation 2 are

$$\ell_{jn} = \Pr(y_j = c_n | \vec{S}_j) = \frac{\Pr(\vec{S}_j | y_j = c_n) \Pr(y_j = c_n)}{\Pr(\vec{S}_j)}$$

$$= \frac{\Pr(y_j = c_n) \Pr(\vec{S}_j | y_j = c_n)}{\sum_{n=1}^{N} \left(\Pr(y_i = c_n) \Pr(\vec{S}_j | y_j = c_n) \right)}$$

$$\ell_{jn} = \frac{f_n \prod_{k=1}^{S} \Pr(s_{jk} | y_j = c_n)^{s_{jk}} \left[1 - \Pr(s_{jk} | y_j = c_n)\right]^{(s_{jk}-1)}}{\sum_{n=1}^{N} \left(f_n \prod_{k=1}^{S} \Pr(s_{jk} | y_j = c_n)^{s_{jk}} \left[1 - \Pr(s_{jk} | y_j = c_n)\right]^{(s_{jk}-1)} \right)} \quad (3)$$

In the last step, we use f_n (from $\vec{F}$) for $\Pr(y_j = c_n)$ which ties the individual cause assignments back to the CSMFs

Computation

- Take sample from joint distribution $(\vec{F}, \vec{Y})$ using a Gibbs sampler

Computation

- Take sample from joint distribution $(\vec{F}, \vec{Y})$ using a Gibbs sampler
- The model defines the conditional distributions we need, proceed by

Computation

- Take sample from joint distribution $(\vec{F}, \vec{Y})$ using a Gibbs sampler
- The model defines the conditional distributions we need, proceed by

    1. Assuming a reasonable starting value for $\vec{F}$, *use equation 2 to draw a cause for each death*

Computation

- Take sample from joint distribution $(\vec{F}, \vec{Y})$ using a Gibbs sampler
- The model defines the conditional distributions we need, proceed by

  1. Assuming a reasonable starting value for $\vec{F}$, *use equation 2 to draw a cause for each death*
  2. *Use these cause assignments in equation 1 to draw a new CSMF vector $\vec{F}$*

InSilicoVA algorithm - equations and computation 4

Computation

- Take sample from joint distribution $(\vec{F}, \vec{Y})$ using a Gibbs sampler
- The model defines the conditional distributions we need, proceed by

  1. Assuming a reasonable starting value for $\vec{F}$, *use equation 2 to draw a cause for each death*
  2. *Use these cause assignments in equation 1 to draw a new CSMF vector $\vec{F}$*
  3. Use this new $\vec{F}$ in equation 2 to draw a new cause for each death, etc ... repeat these two steps *many* times

Computation

- Take sample from joint distribution $(\vec{F}, \vec{Y})$ using a Gibbs sampler
- The model defines the conditional distributions we need, proceed by

  1. Assuming a reasonable starting value for $\vec{F}$, *use equation 2 to draw a cause for each death*
  2. *Use these cause assignments in equation 1 to draw a new CSMF vector $\vec{F}$*
  3. Use this new $\vec{F}$ in equation 2 to draw a new cause for each death, etc . . . repeat these two steps *many* times

- After both $\vec{F}$ and $\vec{Y}$ have settled into a stationary distribution, take a number of samples from a large number of steps and record the values of $\vec{F}$, $\vec{Y}$, $\vec{L}$

InSilicoVA algorithm - equations and computation 4

Computation

- Take sample from joint distribution $(\vec{F}, \vec{Y})$ using a Gibbs sampler
- The model defines the conditional distributions we need, proceed by

  1. Assuming a reasonable starting value for $\vec{F}$, *use equation 2 to draw a cause for each death*
  2. *Use these cause assignments in equation 1 to draw a new CSMF vector $\vec{F}$*
  3. Use this new $\vec{F}$ in equation 2 to draw a new cause for each death, etc ... repeat these two steps *many* times

- After both $\vec{F}$ and $\vec{Y}$ have settled into a stationary distribution, take a number of samples from a large number of steps and record the values of $\vec{F}$, $\vec{Y}$, $\vec{L}$

- These samples approximate the joint distribution $(\vec{F}, \vec{Y})$ and the $\vec{L}$ that go along with them

InSilicoVA algorithm - equations and computation 4

Computation

- Take sample from joint distribution $(\vec{F}, \vec{Y})$ using a Gibbs sampler
- The model defines the conditional distributions we need, proceed by

  1. Assuming a reasonable starting value for $\vec{F}$, *use equation 2 to draw a cause for each death*
  2. *Use these cause assignments in equation 1 to draw a new CSMF vector $\vec{F}$*
  3. Use this new $\vec{F}$ in equation 2 to draw a new cause for each death, etc . . . repeat these two steps *many* times

- After both $\vec{F}$ and $\vec{Y}$ have settled into a stationary distribution, take a number of samples from a large number of steps and record the values of $\vec{F}$, $\vec{Y}$, $\vec{L}$
- These samples approximate the joint distribution $(\vec{F}, \vec{Y})$ and the $\vec{L}$ that go along with them
- Summarize the margins of the sampled $\vec{F}$, $\vec{Y}$ and $\vec{L}$ as necessary to produce desired outputs