A Compositional Approach for Learning High-Dimensional Distributions from Small Samples

Francisco Sánchez-Vega[1], Jason Eisner[2], Donald Geman[1] and Laurent Younes[1].

Applied Mathematics and Statistics Department[1] and Computer Science Department[2].
Johns Hopkins University, Baltimore, MD.

Set of n i.i.d. samples from $P(X_1,\dots,X_d)$, typically $n \ll d$

d variables

n samples

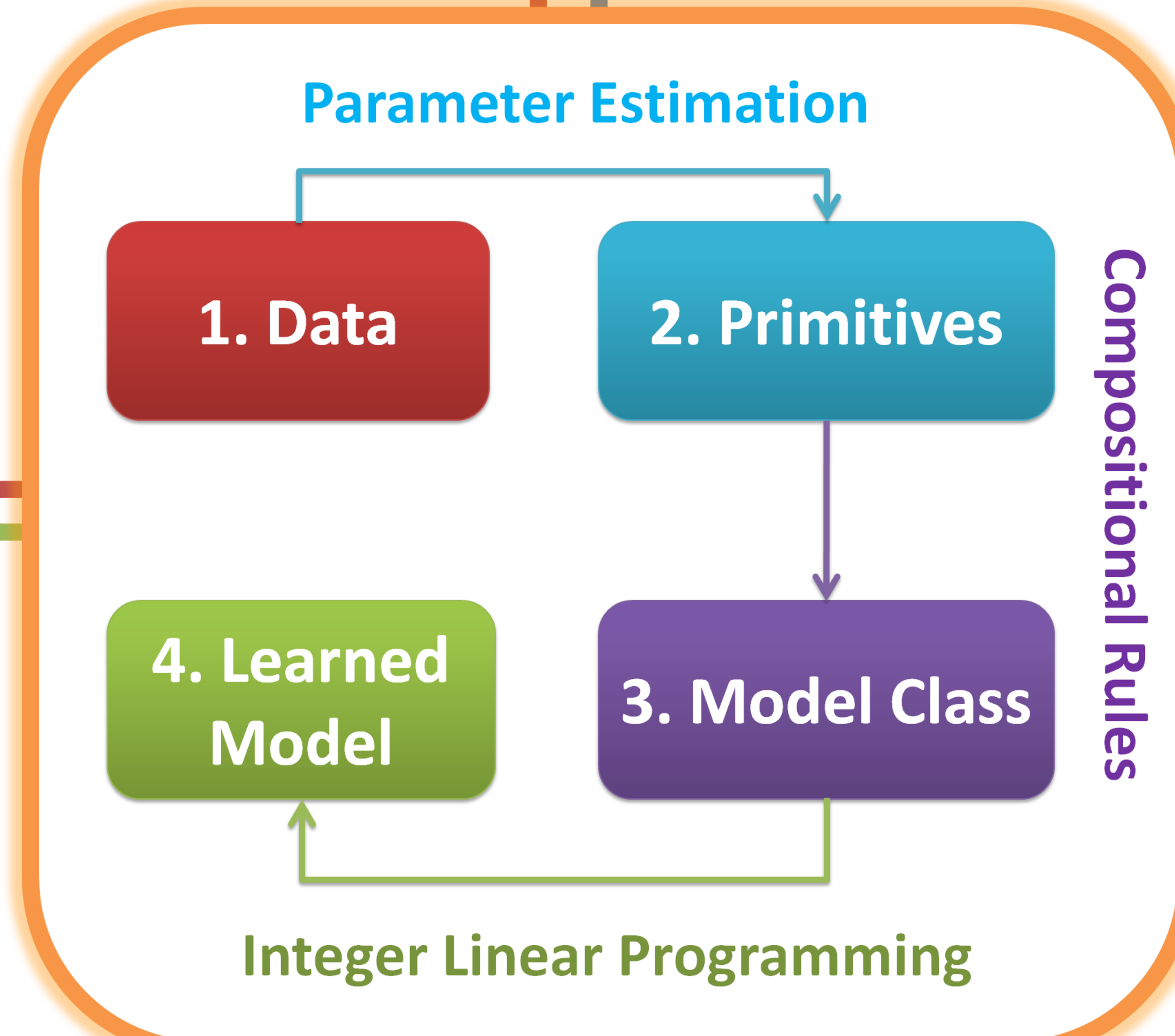
X_1	X_2	…	X_d
$x_1^{(1)}$	$x_2^{(1)}$	…	$x_d^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	…	$x_d^{(2)}$
…	…	…	…
$x_1^{(n)}$	$x_2^{(n)}$	…	$x_d^{(n)}$

Our Method: Competitive Assembly of Marginals

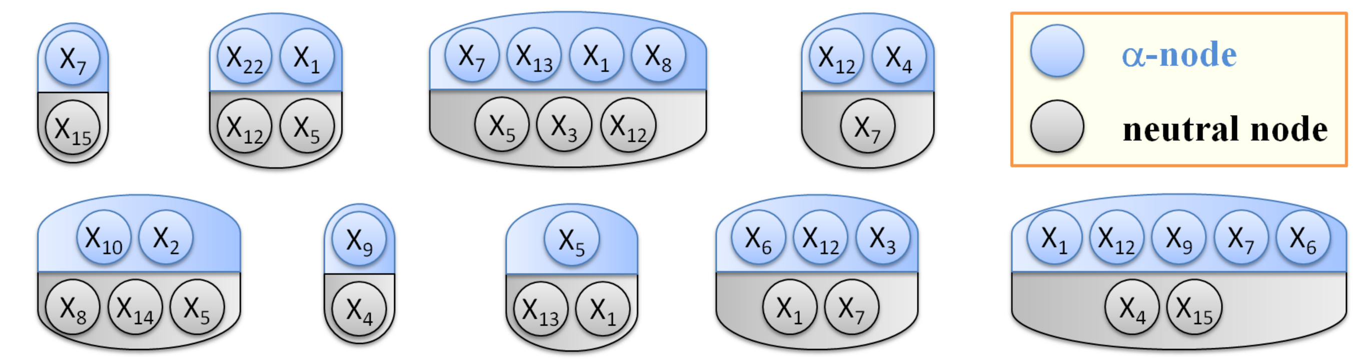
• **Motivation:** Model selection in "small n, large d" settings.
- Need mechanisms to avoid model overfitting.

• **Strategy:** Learn graphical model from data (structure and parameters).
- Adapt model complexity to sample size.
- Enforce biases that restrict the set of admissible distributions.

• Stepwise primitive selection process.
• Each increment of parametric complexity needs to be justified by likelihood gain.

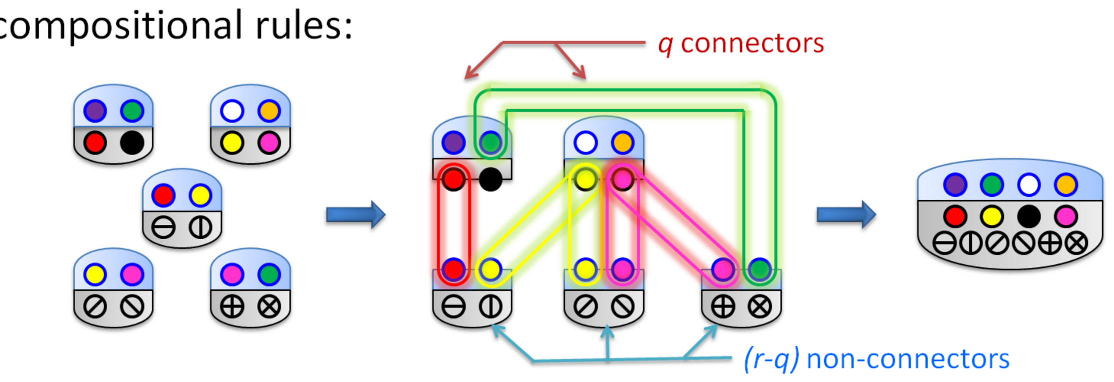


Low-dimensional marginals selected from data (*primitives*)



Primitives are triplets $\phi=(\pi, A, \xi)$, where
- π is a probability distribution on $J(\pi)\subset D$.
- $A \subset J(\pi)$ is the set of **α-nodes**, $A\neq\emptyset$.
- $N=J(\pi)\setminus A$ is the set of **neutral nodes**.
- $\xi: D \to \mathbb{R}$ is a function to control primitive overlap.

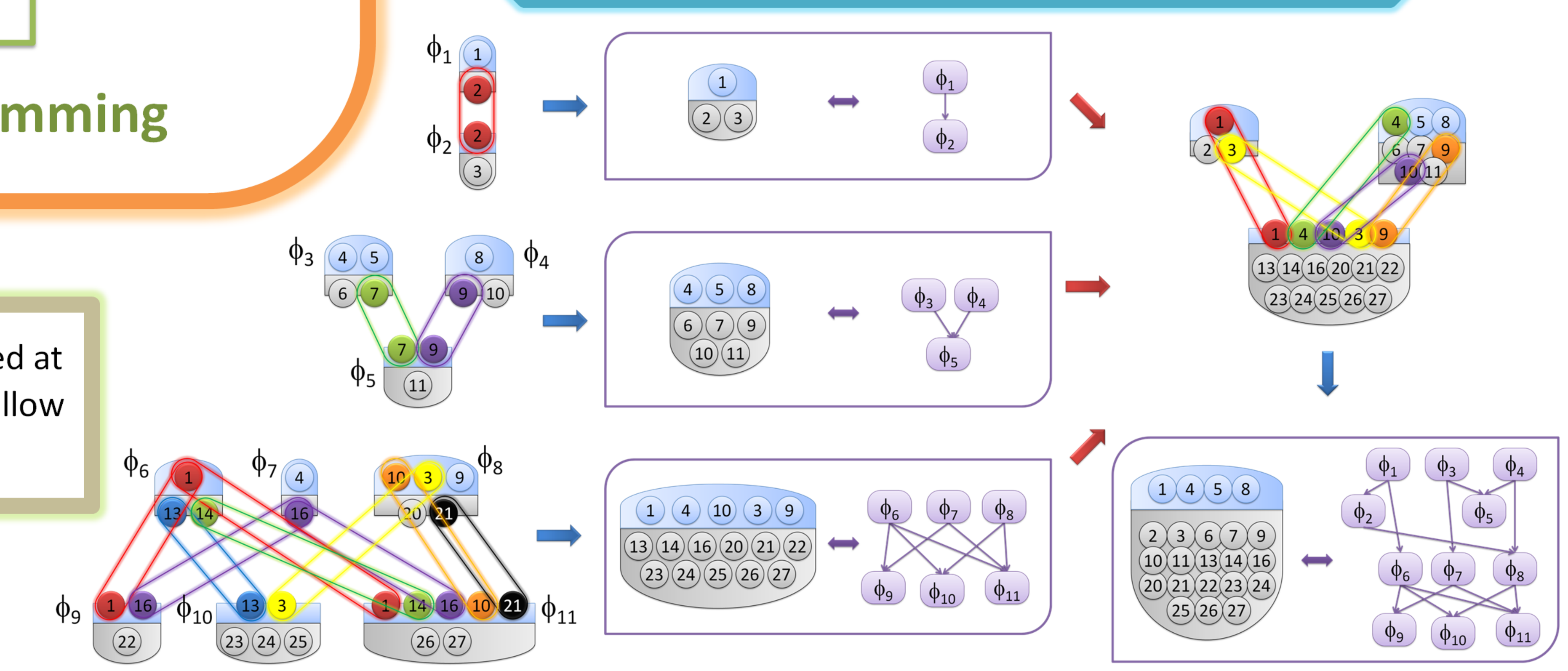
Primitives can be merged into larger distributions, subject to some compositional rules:



The merged distribution is obtained by conditioning on α–nodes:

$$\pi(x_S) = \prod_{k=1}^{q} \pi_k(x_{J(\pi_k)}) \prod_{k=q+1}^{r} \pi_k(x_{J(\pi_k)} \mid x_{A_k})$$

where $\phi_k=(\pi_k, A_k, \xi_k)$ for primitives $(\phi_1,\dots,\phi_r)$ and $S=\bigcup_{k=1}^{r} J(\pi_k)$.

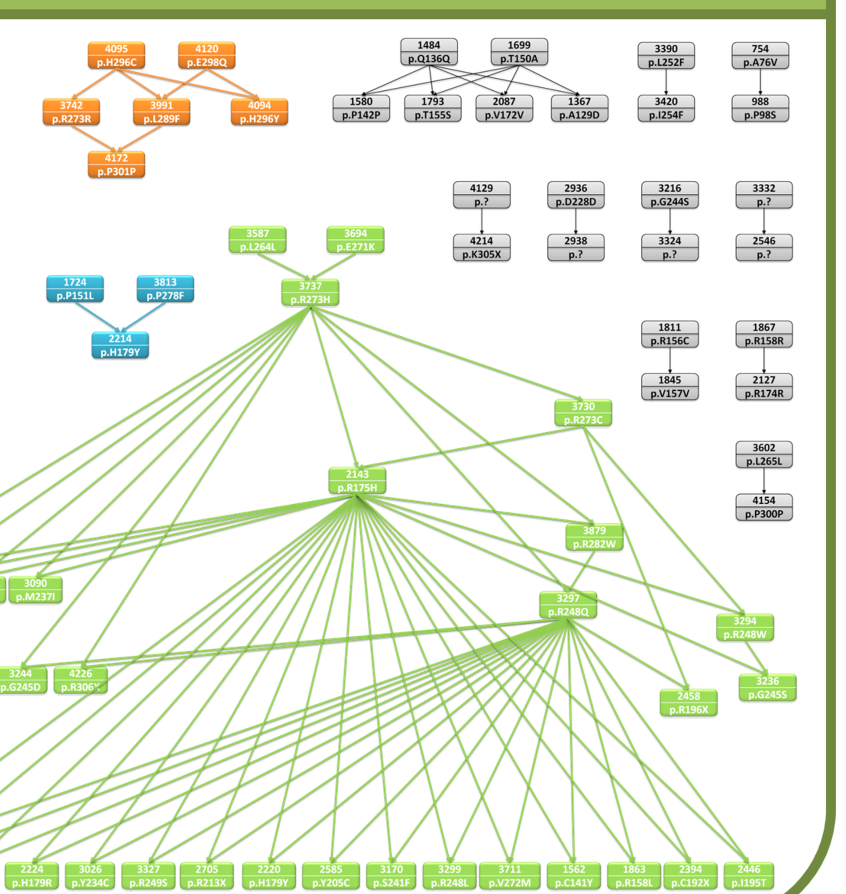


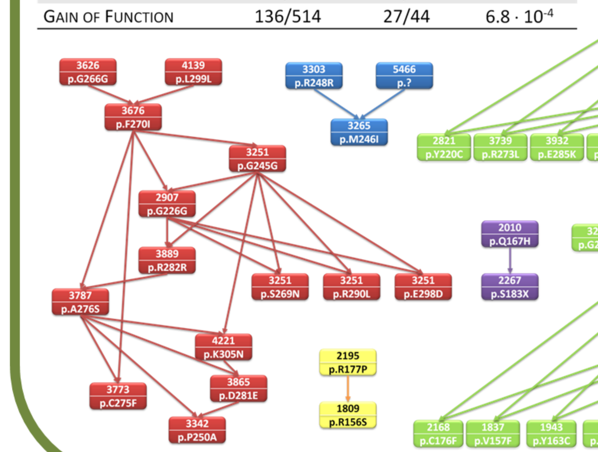
Legal sequence of merges ↔ Directed acyclic graph of primitives

Example: Somatic Mutations in Gene *TP53*

Topography		Morphology
(light green)	[59 terms]	[129 terms]
(red)	Lung	Squamous cell carcinoma, Adenocarcinoma
(orange)	Stomach	
(cyan)		Nodular Melanoma
(blue)		Hemangiosarcoma
(purple)	Thyroid	Papillary Carcinoma, Mature T-Cell Lymphoma
(yellow)		Yolk Sac Tumor, Astrocytoma Anaplastic

Property	All mutations	Network	p-value
CpG Association	213/2000	19/86	0.0012
Average Mutation Rate	0.1381	0.2832	$1.3 \cdot 10^{-4}$
Deleterious (SIFT)	290/490	33/40	$2.58 \cdot 10^{-4}$
Deleterious (AGVGD)	279/490	33/40	$8.6 \cdot 10^{-4}$
Loss of Function	443/514	41/44	0.0402
Gain of Function	136/514	27/44	$6.8 \cdot 10^{-4}$



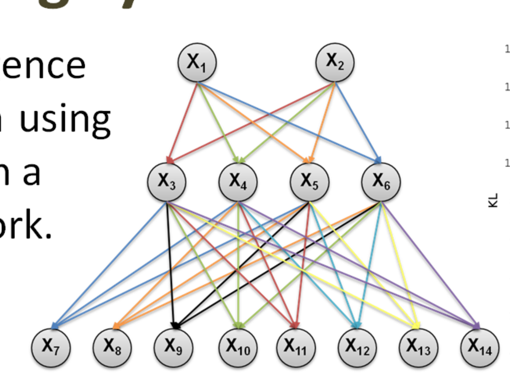


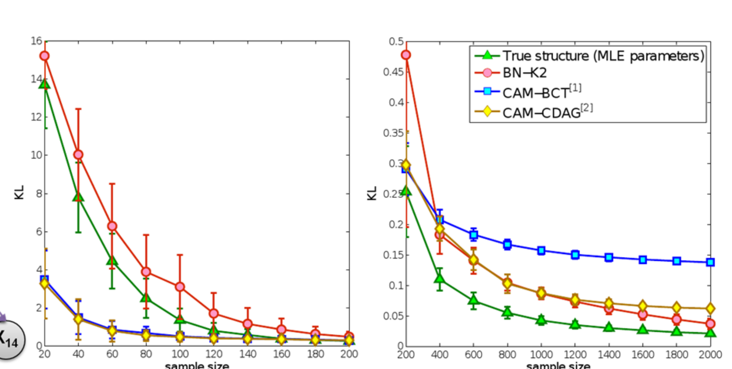
• Besides the local constraints enforced at the primitive and merge levels, we allow for global topological constraints.

Graphical model that estimates $P(X_1,\dots,X_d)$

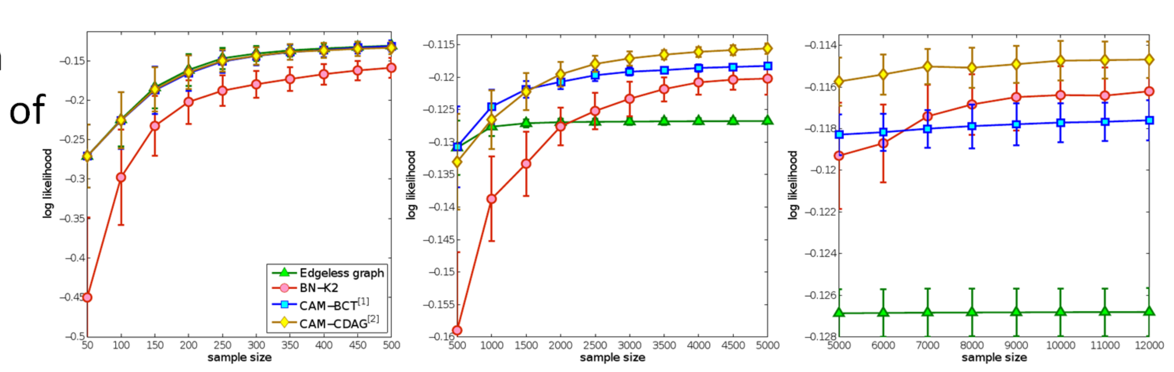
Validation using synthetic and real data

a) Kullback-Leibler divergence to the true distribution using synthetic samples from a known Bayesian network.





b) Predictive performance on randomly selected subsets of holdout samples from the *20newsgroups* dataset.



References

[1] Francisco Sanchez-Vega, Jason Eisner, Laurent Younes, Donald Geman: "Learning Multivariate Distributions by Competitive Assembly of Marginals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, April 2012.

[2] Francisco Sanchez-Vega: "Small Sample Learning of Multivariate Distributions with Compositional Graphical Models," Ph.D. dissertation. The Johns Hopkins University, October 2012.

Please send correspondence to Francisco Sanchez-Vega (*sanchez@cis.jhu.edu*) at the Center for Imaging Sciences, 307C Clark Hall. The Johns Hopkins University. 3400 N. Charles Street, Baltimore, MD, 21218-2686, USA. *http://www.cis.jhu.edu*