Variational inference for mixture models

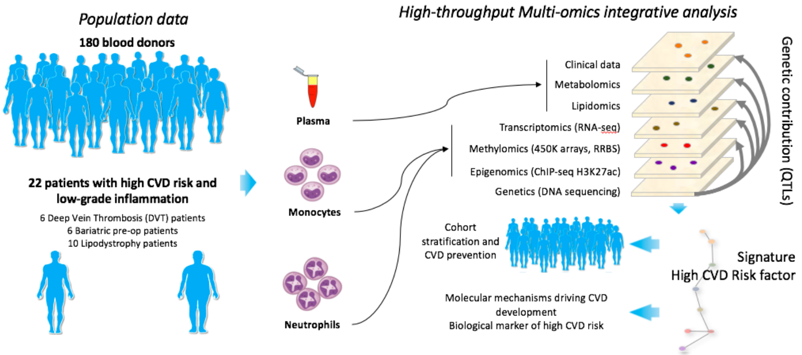
Alessandra Cabassi
Dr Paul D. W. Kirk

24 April 2019

MRC Biostatistics Unit
UNIVERSITY OF CAMBRIDGE

Motivation

Multi-omic analysis of cardiovascular disease (CVD) risk data
with Dr Denis Seyres and Dr Mattia Frontini – Department of Hæmatology



Motivation

CVD risk data	Cell type	Variables	Observations
Epigenomics		25600	172
		26300	128
Methylomics		26214	193
		21442	187
Transcriptomics		11370	203
		24224	198
Lipidomics		123	192
Metabolomics		988	200

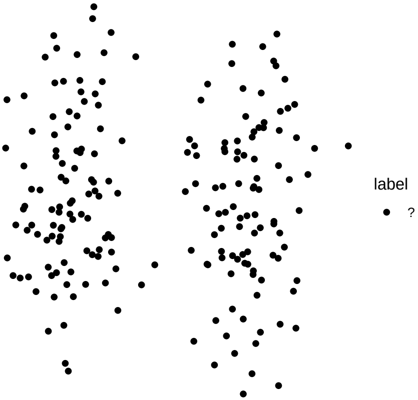
We need:

- Scalable approximate inference method for **mixture models**
- That allows to combine **different types of data**
- And to perform **feature selection**

Motivation

Why feature selection?

Example: Mixture of Gaussians

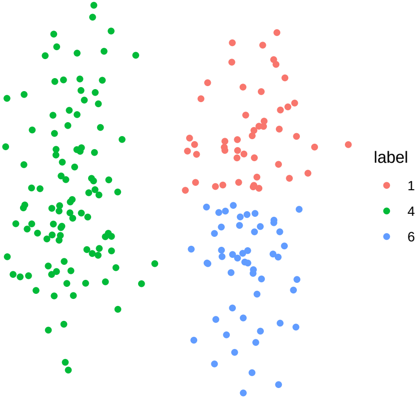


! Noisy features can degrade the performance of most learning algorithms Law et al. (2004)

Motivation

Why feature selection?

Example: Mixture of Gaussians fitted using *both features*



! Noisy features can degrade the performance of most learning algorithms Law et al. (2004)

Motivation

Why feature selection?

Example: Mixture of Gaussians fitted using *only the relevant feature*

! Noisy features can degrade the performance of most learning algorithms Law et al. (2004)

Approximate inference

Why?

Given a joint model for our hidden variables z and observed variables x, $p(x, z)$, inference about the unknown is through the posterior

$$p(z|x) = \frac{p(z, x)}{p(x)}$$

For most interesting models, the denominator is not tractable, so we appeal to approximate posterior inference

Approximate inference

Which type?

Stochastic approximations $\rightarrow$ Sampling

+ Asymptotically exact
+ Easily applicable general-purpose algorithms
– Computationally expensive
– Storage intensive

Deterministic approximations $\rightarrow$ Structural assumptions

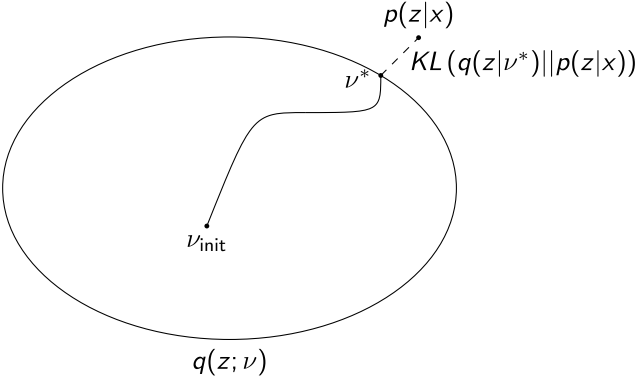
+ Computationally efficient
+ Efficient representation
– Often hard work to derive
– Not guaranteed to converge to global optimum

Brodersen (2010)

Variational inference

Main idea

Posit a variational family of distributions over the latent variables $q(z; \nu)$ and fit the variational parameters ν to be close (in Kullback-Leibler divergence) to the exact posterior



Variational inference

History

Variational inference (VI) adapts **ideas from statistical physics** to probabilistic inference

Variational inference

History

Variational inference (VI) adapts **ideas from statistical physics** to probabilistic inference

🇺🇸 1980s: Peterson and Anderson (1987) used **mean-field methods to fit a neural network**

Variational inference

History

Variational inference (VI) adapts **ideas from statistical physics** to probabilistic inference

🇺🇸 1980s: Peterson and Anderson (1987) used **mean-field methods to fit a neural network**
Early 1990s: This idea was picked up by Jordan's lab who **generalised it to many probabilistic models** (a review paper is Jordan, Ghahramani, Jaakkola and Saul, 1999)

Variational inference

History

Variational inference (VI) adapts **ideas from statistical physics** to probabilistic inference

🇺🇸 1980s: Peterson and Anderson (1987) used **mean-field methods to fit a neural network**
Early 1990s: This idea was picked up by Jordan's lab who **generalised it to many probabilistic models** (a review paper is Jordan, Ghahramani, Jaakkola and Saul, 1999)

🇨🇦 In parallel: Hinton and Van Camp (1993) developed mean-field for neural networks. Neal and Hinton (1993) **connected this idea to the EM algorithm**, which lead to further **variational methods for mixtures of experts** (Waterhouse et al., 1996)

Variational inference

Preliminary definitions

Entropy: (information theory) average rate at which information is produced by a stochastic source of data

Given a random variable x with probability density function $p(x)$

$$H(x) = -\int p(x) \log p(x) dx = \mathbb{E}_p[-\log p(x)]$$

Entropy increases as the distribution becomes broader

Variational inference

Preliminary definitions

Kullback-Leibler divergence (relative entropy): measure of how one probability distribution is different from a second, reference probability distribution

$p(z)$: unknown distribution
$q(z)$: approximating distribution

$$
\begin{aligned}
\mathrm{KL}(q||p) &= -\int q(z) \log \left\{ \frac{p(z)}{q(z)} \right\} dz \\
&= -\int q(z) \log p(z) dz - \underbrace{\left(-\int q(z) \log q(z) dz \right)}_{\text{Entropy of q}}
\end{aligned}
$$

Properties:

- $\mathrm{KL}(q||p) \geq 0$
- $\mathrm{KL}(q||p) = 0$ iff $p = q$
- $\mathrm{KL}(q||p) \neq \mathrm{KL}(p||q)$

Variational inference

The evidence lower bound (ELBO)

Recall

$$p(z|x) = \frac{p(x,z)}{p(x)}$$

$$\log p(x,z) = \log\left[p(z|x)p(x)\right]$$

$$\int \mathbf{log}\frac{\boldsymbol{p(x,z)}}{q(z)}q(z)dz \;\; \mathbf{=} \int \mathbf{log}\frac{\boldsymbol{p(x)p(x|z)}}{q(z)}q(z)dz$$

$$\int \log \frac{p(x,z)}{q(z)}q(z)dz = \log p(x) - \left[-\int \log \frac{p(z|x)}{q(z)}q(z)dz\right]$$

$$\underbrace{\mathcal{L}(q)}_{\text{ELBO}} = \log p(x) - \underbrace{KL(q||p)}_{\text{KL divergence}}$$

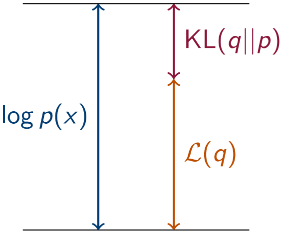
Variational inference

The evidence lower bound (ELBO)

$$\log p(x) = \underbrace{\mathcal{L}(q)}_{\text{ELBO}} + \underbrace{KL(q||p)}_{\text{KL divergence}}$$

$$\mathrm{KL}(q||p) \geq 0$$

$$\log p(x) \geq \mathcal{L}(q)$$



KL is intractable, so we optimise the ELBO instead

Variational inference

The evidence lower bound (ELBO)

$$\begin{aligned}\mathcal{L}(q) &= \int \log p(x,z)q(z)dz - \int \log q(z)q(z)dz \\ &= \underbrace{\mathbb{E}_q[\log p(x,z)]}_{\substack{(1) \\ \text{Exp [log prior +} \\ \text{log likelihood]}}} + \underbrace{\mathbb{E}_q[-\log q(z)]}_{\substack{(2) \\ \text{Entropy of q}}}\end{aligned}$$

The ELBO trades off two terms:

(1) Prefers q to place its mass on the maximum a posteriori estimate

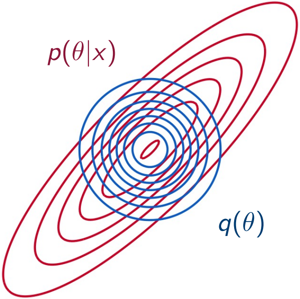
(2) Encourages q to be diffuse

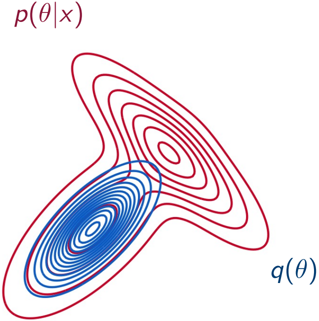
! The ELBO is not convex

Variational inference

Some properties

Bishop (2006)





$q(\theta)$ tends to be 0 where $p(\theta|x)$ is 0.

VI may lead to a local minimum.

Variational inference

Mean-field approximation

We need to specify the form of q(z). The mean-field family is fully factorised:

$$q(z) = \prod_{i=1}^{M} q_i(z_i)$$

Optimise the ELBO. Traditionally, VI uses coordinate ascent:

$$\log q_i^*(z_i) \propto \mathbb{E}_{j \neq i}[\log p(x, z)]$$

Iteratively update each parameter, holding others fixed.

Variational inference

Coordinate ascent (CAVI) algorithm

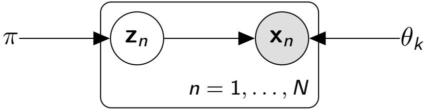
Input : A model $p(X, \theta)$, a dataset X
Output : A variational density $q(\theta) = \prod_j q_j(\theta_j)$
Initialise: Variational factors $q_j(\theta_j)$
do
  for $j \in \{1, \ldots, J\}$ **do**
    set $q_j(\theta_j) \propto \exp\{\mathbb{E}_{i \neq j}[\log p(X, \theta)]\}$
  end
  compute the ELBO $\mathcal{L}(q)$
while *the ELBO has not converged*;
return $q(\theta)$.

Mixture models

Main idea

$$p(x) = \sum_{k=1}^{K} \pi_k f_x(x|\theta_k).$$

$f_{\boldsymbol{x}}$ parametric density that depends on the parameter(s) θ_k
π_k cluster weights



Example: Mixture of Gaussians

$$x_n \sim \prod_k \mathcal{N}(\mu_k, \Lambda_k^{-1})^{z_{nk}}$$

$$z_{nk} \sim \text{Bernoulli}(\pi_k)$$

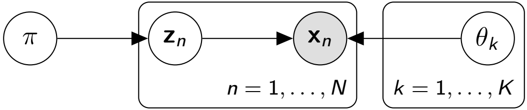
Input : A model $p(x, z|\theta, \pi)$, a dataset X
Output : The parameters θ^*, π^* maximising the log-likelihood
Initialise: Parameters π, θ, responsibilities $\mathbb{E}[z_{nk}]$
do
  Expectation step: evaluate the responsibilities $\mathbb{E}[z_{nk}]$
  Maximisation step: update the other parameters in turn
while *convergence is not reached*;
return θ^*, π^*

Mixture models

...in the Bayesian framework

$$p(x) = \sum_{k=1}^{K} \pi_k f_x(x|\theta_k).$$

$f_{\boldsymbol{x}}$ parametric density that depends on the parameter(s) θ_k
π_k cluster weights



Example: Mixture of Gaussians

$$\pi \sim \text{Dirichlet}(\alpha_0, \ldots, \alpha_0)$$

$$\theta = \{\mu, \Sigma\}$$

$$\mu_k \sim \mathcal{N}(m_0, (\beta_0 \Lambda_k)^{-1})$$

$$\Lambda_k \sim \mathcal{W}(W_0, \nu_0)$$

Variational inference for mixture models

Approximate the true posterior with a variational distribution q

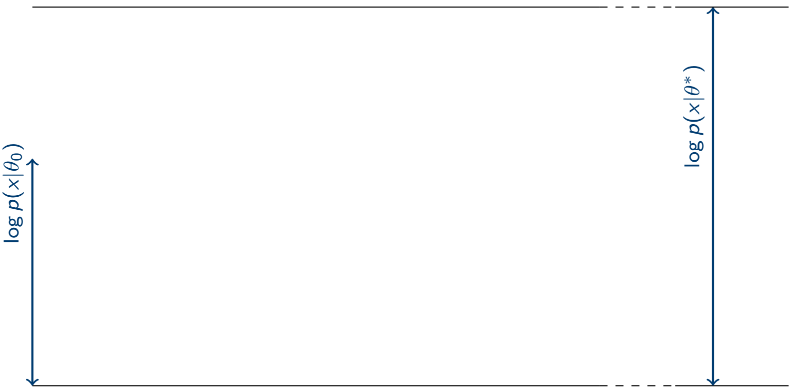
$$q(z, \theta, \pi) = q(z)q(\theta, \pi)$$

EM-type algorithm

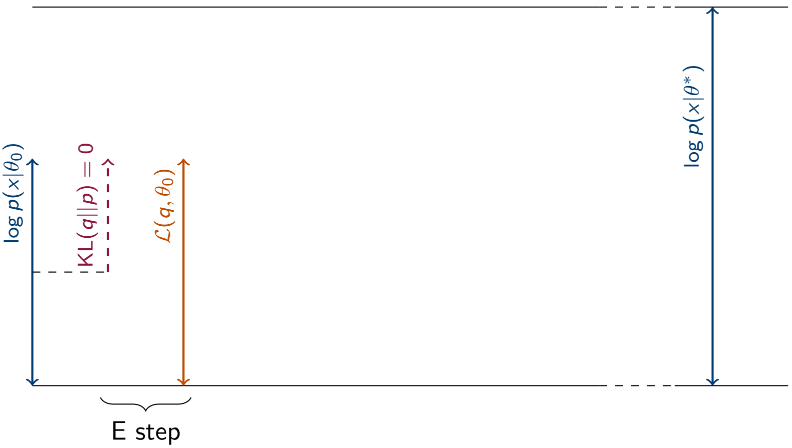
Input : A model $p(x, z, \pi, \theta)$, a dataset X
Output : A variational density $q(z^*, \pi^*, \theta^*) = q(z^*)q(\pi^*, \theta^*)$
Initialise: Parameters π, θ, responsibilities $\mathbb{E}[z_{nk}]$
do
  Expectation step: evaluate the responsibilities $\mathbb{E}[z_{nk}]$
  Maximisation step: update the other hyperparameters in turn
while *the ELBO has not converged*;
return $q(z^*, \pi^*, \theta^*)$

Bishop (2006)

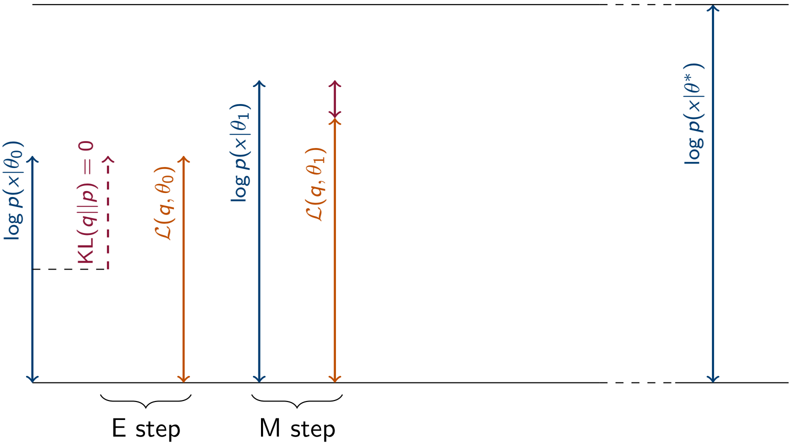
Variational inference for mixture models



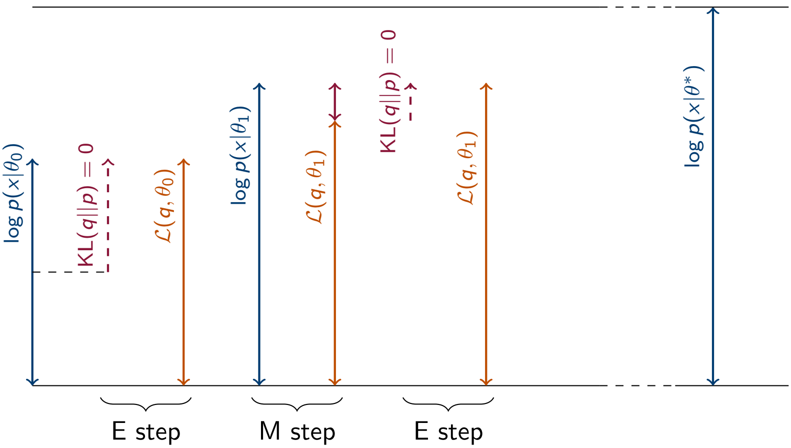
Variational inference for mixture models



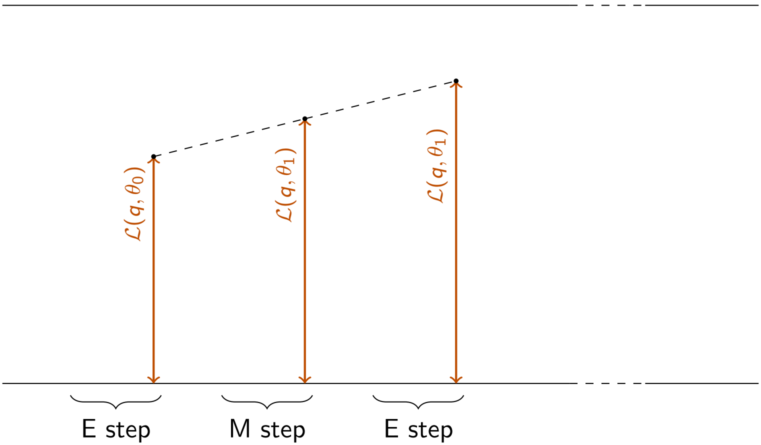
Variational inference for mixture models



Variational inference for mixture models



Variational inference for mixture models



Lower bound used to check:

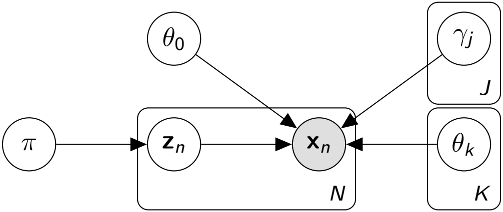
- Correctness of update equations
- Convergence

Mixture models

Feature selection

$$p(x) = \sum_k \pi_k \prod_j p_{x_j}(x_j|\theta_k)^{\gamma_j} p_{x_j}(x_j|\theta_0)^{1-\gamma_j}$$

$$\gamma_j \sim \text{Bernoulli}(\delta_j)$$



Our project

Plan

Implementation and analysis of the following mixtures using VI:

	Basic model	Feature selection	Model selection
Gaussian	$\checkmark$	$\sim$	$\sim$
Categorical	$\checkmark$		
Gaussian + Categorical			

$\checkmark$	already studied in the literature, code available online (Bishop 2006, Ahlmann-Eltze and Yau 2018)
$\sim$	already studied in the literature, code *not* available online (Constantinopoulos et al. 2006)

Our project

Current status

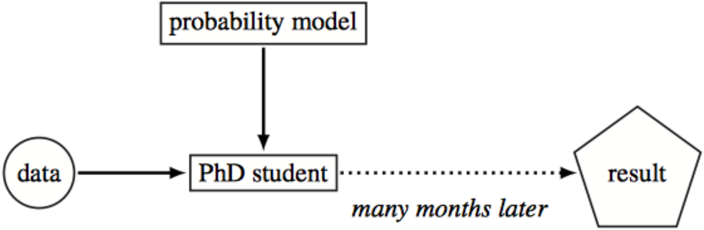


Figure 1: The "how hard could it be?"™ way of probabilistic modeling.

Kucukelbir (2015)

Our project

Current status

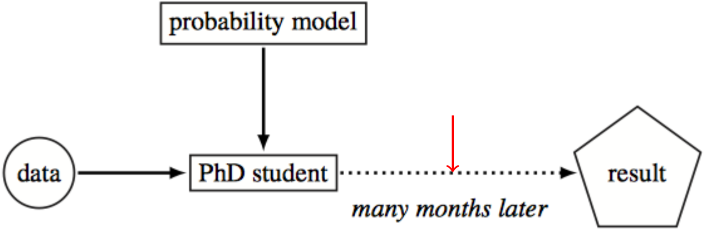


Figure 1: The "how hard could it be?"™ way of probabilistic modeling.

Kucukelbir (2015)

Our project

Future work

- Complete R package "vimix" `https://acabassi.github.io/vimix/`
- Apply to CVD risk data
- Explore automated tools?
  E.g. TensorFlow Probability, PyMC3, Edward, Stan

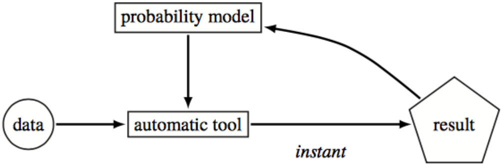


Figure 3: The probabilistic programming way of probabilistic modeling.

Thanks for listening!

MRC | Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

alessandra.cabassi@mrc-bsu.cam.ac.uk
alessandracabassi.com
@SandyCabassi

References I

Ahlmann-Eltze, C. and Yau, C., 2018.
MixDir: Scalable Bayesian Clustering for High-Dimensional Categorical Data.
In IEEE 5th International Conference on Data Science and Advanced Analytics.

Bishop, C.M., 2006.
Pattern recognition and machine learning.
Springer, 128.

Blei, D.M., Kucukelbir, A. and McAuliffe, J.D., 2017.
Variational inference: A review for statisticians.
Journal of the American Statistical Association, 112(518), pp.859-877.

Corduneanu, A. and Bishop, C.M., 2001.
Variational Bayesian model selection for mixture distributions.
Artificial intelligence and Statistics, vol. 2001, pp. 27-34.

Constantinopoulos, C., Titsias, M.K. and Likas, A., 2006.
Bayesian feature and model selection for Gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(6), pp.1013-1018.*

Guan, Y., Dy*, J.G., Niu, D. and Ghahramani, Z., 2010.
Variational inference for nonparametric multiple clustering.
MultiClust Workshop, KDD-2010.

* The only woman in this bibliography!

References II

Hinton, G. and Van Camp, D., 1993.
Keeping neural networks simple by minimizing the description length of the weights.
In in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S. and Saul, L.K., 1999.
An introduction to variational methods for graphical models.
Machine learning, 37(2), pp.183-233.

Kucukelbir, A., 2015.
Probabilistic Modeling in Stan.
Notes

Law, M.H., Figueiredo, M.A. and Jain, A.K., 2004.
Simultaneous feature selection and clustering using mixture models.
IEEE transactions on pattern analysis and machine intelligence, 26(9), pp.1154-1166.

Neal, R.M. and Hinton, G.E., 1998.
A view of the EM algorithm that justifies incremental, sparse, and other variants.
In Learning in graphical models (pp. 355-368). Springer, Dordrecht.

Peterson, C., and Anderson, J.R., 1987.
A mean field theory learning algorithm for neural networks.
Complex Systems, 1, pp.995-1019.

Figures credits

Slide 1
Denis Seyres

Slide 2
Mikael Häggström and A. Rad
Image:Hematopoiesis (human) diagram.png by A. Rad, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=7351905