A general-purpose Bayesian model for estimating election results in Mexico

Gian Carlo Di-Luvi
Department of Statistics
University of British Columbia
gian.diluvi@stat.ubc.ca

Abstract

In Mexico, estimating the results of the election the day it takes place is a tradition that gives certainty to the electoral process. For this purpose, electoral authorities organize a quick count, in which a random sample of polling stations is used to estimate the results of the election. Multiple Bayesian models have been used in Mexican quick counts since 2006. However, these models assume that candidates are independent, which has a negative impact on model accuracy. In this work, we propose a general-purpose model that takes care of these issues by forcing the proportions of votes to reside in the probabilistic simplex and assuming an arbitrary covariance structure. We discuss uninformative prior distributions for both the proportions in the simplex and the covariance matrix and compare the model with one of the models used in the 2018 Mexican presidential election.

1 Introduction

In Mexico, presidential and gubernatorial elections take place every six years. Mexico has a multi-party electoral system where the candidate that receives the largest amount of votes wins the election. Because election results take a week to be certified, in recent years there has been a widespread effort by government authorities to determine the winner of each election the day it takes place. For this purpose, a sample of polling stations is usually selected before the election and statistical estimates based on those polling stations are computed as the results become available on election day—a process called a *quick count* (Carrera Barroso, 2019).

Multiple Bayesian models have been used in Mexico's quick counts. Mendoza and Nieto-Barajas (2016) proposed a simple Bayesian model with a tractable posterior to estimate the proportion of votes in favor of each candidate in the 2006 and 2012 elections. Cerrillo and Barajas (2019) used the same model in the 2016 elections of Veracruz, a state in the east cost of Mexico. Di-Luvi et al. (2018) modified that same model and used it to predict the results of the 2018 election. Anzarut et al. (2018) proposed a Bayesian multilevel regression model and used it in the 2018 election as well.

On election day, data from the polling stations arrive as field personnel finish counting the votes: every 5 minutes, new data are supplied and models have to be retrained. Thus, models should be rich enough to do accurate inference with small sample sizes while allowing for fast online inference. The model in (Anzarut et al., 2018) accounts for data missing not at random but is computationally expensive—for the presidential election, it took well over 5 minutes to run even when parallelising over 48 CPU cores. On the other hand, although the model proposed by (Mendoza and Nieto-Barajas, 2016) is easy to train, they rely on a tractable posterior distribution that severly overestimates the voter turnout. Di-Luvi et al.

(2018) modified the prior distribution to remedy this, but they did not account for the corresponding change in the posterior.

One of the main issues with previous models is that the candidates are treated independently. In this work, we address this issue by forcing the proportions of votes to reside in the probabilistic simplex Δ^{J-1}. Furthermore, by working with a multivariate joint model for all candidates, we design models with arbitrary covariance structures. Finally, we implement an MCMC sampling scheme in Stan Stan Development Team (2020) to fit all these models.

In Section 2, we present the original model in Mendoza and Nieto-Barajas (2016), as well as the modifications by Di-Luvi et al. (2018). In Section 3, we show how to modify the model to properly account for both the proportions of votes and an arbitrary covariance structure. We implement the original models and two new models in Section 4, and conclude the paper in Section 5.

2 Background

Quick counts are organized by the Mexican Electoral Institute but carried out by a committee of statisticians. Each member of the committe proposes and implements a different model. On election day, each model is trained every 5 minutes when new information becomes available, and the results from all models are combined into a single interval per candidate. When the committee decides that the sample is sufficiently large, the final combined intervals are reported to the government authorities.

To determine the sample of polling stations, a stratified sampling design is developed by the committee. The observational unit corresponds to the polling station and the strata—which vary from election to election—are defined based on geographical variables. Suppose that the population of polling stations is divided into N strata, and that stratum i has K_i polling stations. Although the number of voters is unknown before the election, the number of *registered voters* per polling station is known in advance. Let n_i^k be the number of registered voters in polling station k of stratum i.

The observables in the model correspond to the number of votes, which for each candidate j, stratum i, and polling station k we denote by X_{ij}^k. The latent variables correspond to the proportion of votes, which for candidate j and in stratum i are denoted by θ_{ij}. In these definitions, we use the term candidate loosely to refer to actual candidates as well as the set of null votes and the set of votes that were not cast. Note that we assume that all polling stations within a stratum have the same proportion of votes per candidate, i.e.

$$\theta_{ij} = \frac{\sum_{k=1}^{K_i} X_{ij}^k}{\sum_{k=1}^{K_i} n_i^k}$$

does not depend on k.[1] The national proportion of votes in favor of candidate j is then

$$\theta_j = \frac{1}{n} \sum_{i=1}^{N} n_i \theta_{ij}, \tag{1}$$

where $n = \sum_{i=1}^{N} n_i$ is the total number of nation-wide registered voters. Observe that θ_j refers to the proportion of votes out of registered voters, rather than out of cast votes. We thus define, for each candidate j, the effective proportion of votes as

$$\lambda_j = \frac{\theta_j}{\sum_{r=1}^{J-1} \theta_r}, \tag{2}$$

[1] It is possible to assume that the proportions of votes vary by polling station, for example using Dirichlet processes (see Lo, 1984; Barajas, 2018), but learning the parameters becomes computationally expensive.

where the sum is over all proper candidates (i.e. excluding θ_J, which corresponds to uncast ballots).

Mendoza and Nieto-Barajas (2016) propose modeling the number of votes in polling station k of stratum i and in favor of candidate j via

$$X_{ij}^k \sim \mathcal{N}\left(n_i^k \theta_{ij}, \frac{\tau_{ij}}{n_i^k}\right), \tag{3}$$

where τ_{ij} is a precision parameter.[2] The number of votes are assumed to be independent between candidates, which is inaccurate because, at each polling station, the total votes should add up to the known number of registered voters, n_i^k. Mendoza and Nieto-Barajas (2016) acknowledge this issue but argue that including a correlation structure, although correct in theory, produces small improvements in accuracy.

Because the quick counts are organized by governmental authorities, prior distributions should assign the same likelihood of winning to each candidate. Mendoza and Nieto-Barajas (2016) use uniform distributions for θ_{ij} and an independent improper prior for τ_{ij}. With these, the posterior distribution after observing the information from c_i polling stations in stratum i has a closed form, namely

$$\begin{aligned} p(\theta_{ij}, \tau_{ij} \mid X_{ij}^1, ..., X_{ij}^{c_i}) \propto \mathcal{N}\left(\theta_{ij} \;\middle|\; \frac{\sum_{k=1}^{c_i} x_{ij}^k}{\sum_{k=1}^{c_i} n_i^k},\ \tau_{ij} \sum_{k=1}^{c_i} n_i^k\right) \mathbb{1}(0 < \theta_{ij} < 1) \\ \times \mathsf{Gamma}\left(\tau_{ij} \;\middle|\; \frac{c_i - 1}{2},\ \frac{1}{2}\left\{\sum_{k=1}^{c_i} \frac{(x_{ij}^k)^2}{n_i^k} - \frac{\left(\sum_{k=1}^{c_i} x_{ij}^k\right)^2}{\sum_{k=1}^{c_i} n_i^k}\right\}\right). \end{aligned} \tag{4}$$

Sampling from Eqn. 4 can be done expeditiously for each stratum—there is no need to run an MCMC scheme. Samples from all strata are then combined via Eqns. 1 and 2 to obtain the national estimates for each candidate.

Because data arrive as field officers finish counting the votes, oftentimes entire strata in the sample have no information available. Notice that Eqn. 4 requires $c_i > 2$, i.e. information from at least two polling stations. In those strata, Mendoza and Nieto-Barajas (2016) sample from the prior distribution—which is the probability distribution that describes the uncertainty in the latent variables. However, because all candidates are assumed independent and $(\theta_{i1}, ..., \theta_{iJ})^\top$ is not constrained to add up to 1, it is possible that the estimated voter turnout, defined as

$$\rho = \sum_{j=1}^{J-1} \theta_j,$$

exceeds 100% in some strata—thereby causing the aggregated national average to be overestimated.

To remedy this, Di-Luvi et al. (2018) proposed new prior distributions for θ_{ij}—namely, Beta distributions with a mean equal to the historic voter turnout divided by the number of proper candidates. Although this seemingly takes care of the issue in practice, it does not technically resolve it because it fails to acknowledge that the proportion of votes, as a vector, lives in the probabilistic simplex. Furthermore, Di-Luvi et al. (2018) only use the modified prior when no information is available from a given stratum. If information is available, samples are obtained from the tractable posterior, Eqn. 4, which does not correspond to the correct posterior associated with the modified prior distributions.

[2] Although $\frac{\tau_{ij}}{n_i^k}$ corresponds to the variance of the distribution.

3 Modeling proportions in the simplex

To design a Bayesian model that correctly addresses the estimation of voter turnout, we first write the likelihood of votes in each stratum in Eqn. 3 as a multivariate Normal distribution. Specifically, letting $X_i^k = (X_{i1}^k, ..., X_{iJ}^k)^\top$ and $\theta_i = (\theta_{i1}, ..., \theta_{iJ})^\top$, Eqn. 3 is equivalent to

$$X_i^k \sim \mathcal{N}_J\left(n_i^k \theta_i, \Sigma_i^k\right),$$

where $\Sigma_i^k = \text{diag}(\tau_{i1}, ..., \tau_{iJ})/n_i^k$. Observe that the number of votes is discrete. A multinomial distribution would be technically better suited to model X_i^k, and it would also ensure that the votes add up to the number of registered voters. However, it would also limit the flexibility of the model because, while possible to control the mean, the correlation structure of a multinomial distribution is determined by the means of each marginal Binomial distribution.

Given that the number of registered voters is usually high,[3] using a Normal approximation to a Binomial distribution is well justified. Furthermore, by adding variance parameters independent of the mean, Mendoza and Nieto-Barajas (2016); Di-Luvi et al. (2018) are able to model more complex variability structures. This approach has two shortcomings. The first is that it assumes that the vector θ_i lives in $[0,1]^J$, when in reality θ_i lives in the probabilistic simplex $\Delta^{J-1} = \{\theta \in \mathbb{R}^J \,|\, \theta_j > 0, \sum_{j=1}^J \theta_j = 1\}$. This is the cause of the voter turnout overestimation. The second issue is that it assumes a trivial covariance structure.

We overcome these issues by *(i)* forcing $\theta_i \in \Delta^{J-1}$ and *(ii)* letting Σ_i^k be an arbitrary covariance matrix. The only additional requirements to fit this more complex model are to carefully design a prior distribution for θ_i in Δ^{J-1} and a prior distribution for Σ_i^k in the space of covariance matrices.

3.1 Prior distributions

To design a prior distribution for θ_i, observe that the dimension of Δ^{J-1} is $J-1$; indeed, Δ^{J-1} and $\mathbb{R}^{J-1}$ are isomorphic. Consider the map $\mathcal{A} : \Delta^{J-1} \to \mathbb{R}^{J-1}$ given by

$$\mathcal{A}(\theta_1, ..., \theta_J) = \left(\log \frac{\theta_1}{\theta_J}, ..., \log \frac{\theta_{J-1}}{\theta_J}\right)^\top.$$

$\mathcal{A}$ is known as the *additive log-ratio* (Aitchison, 1982), and its inverse is given by

$$\mathcal{A}^{-1}(x_1, ..., x_{J-1}) = \left(\frac{\exp(x_1)}{1 + \sum_{i=1}^{J-1} \exp(x_i)}, ..., \frac{\exp(x_{J-1})}{1 + \sum_{i=1}^{J-1} \exp(x_i)}, \frac{1}{1 + \sum_{i=1}^{J-1} \exp(x_i)}\right)^\top.$$

The additive log-ratio transform is closely related to the center log-ratio transform, whose inverse is the popular softmax function. However, we choose to work with $\mathcal{A}$ because it is easier to interpret, and thus to define a prior on the transformed values.

In the transformed space $\mathbb{R}^{J-1}$, the coordinate jth corresponds, up to additive constants, to the log of the proportion of votes in favor of candidate j, for $j = 1, ..., J-1$. We assume that the Jth candidate corresponds to the set of uncast ballots, which is of less interest. For the rest of the candidates, we define a Normal prior and choose the mean so that, when mapped back to Δ^{J-1} via $\mathcal{A}^{-1}$, it coincides with that of Di-Luvi et al. (2018): the historic voter turnout divided by the number of candidates. A dependence structure is introduced when mapping back into the simplex via $\mathcal{A}^{-1}$.

Perhaps the best known distribution in the space of covariance matrices is the Wishart distribution, which is the multivariate generalization of the Gamma distribution. The Wishart distribution was popularized because it is the conjugate prior for the covariance matrix of Normally-distributed populations. However, the purpose of the prior being to

[3] In 2018, the average n_i^k was 550.

quantify the prior knowledge of the parameters, we consider it best to choose a distribution that reflects the vague knowledge we have of the correlation structure between candidates.

For this purpose, we decompose $\Sigma_i^k = \text{diag}(\sigma_1, ..., \sigma_J) P \, \text{diag}(\sigma_1, ..., \sigma_J)/n_i^k$, where P is the correlation matrix and σ_j is the variance of each candidate. Lewandowski et al. (2009) propose a method to generate correlation matrices uniformly over the space of all correlation matrices. We use their method, through its STAN implementation, and couple it with uninformative priors for the variances. This results in an uninformative prior for the covariance matrix.

4 Experiments

We study the results of the 2018 Presidential election and compare four different models:

1. The model used by Di-Luvi et al. (2018) in the 2018 elections, which contains modified Beta priors but samples from the original posterior in Mendoza and Nieto-Barajas (2016);
2. that same model but now sampling from the correct posterior via MCMC;
3. the model proposed in Section 3 but assuming a trivial correlation structure, i.e., $P = I_J$;
4. and the same new model but with an arbitrary covariance structure and the uninformative prior on the covariance matrix.

In the third model, the proportions of votes are forced to live in the probabilistic simplex but the correlation structure assumes a simple form, which results in faster sampling.

For model 1, we generate 10,000 observations from the posterior distribution of $(\lambda_1, ..., \lambda_5)$ and ρ. For models 2 and 3, we generate 1,000 observations, and for model 4 only 500 due to the excessive computing time required to fit the model. Code can be found on `https://github.com/GiankDiluvi/bayes-quick-counts-2020`.

We focus our attention on the winner of the election, AMLO. Figure 1 shows the posterior distribution of the proportion of votes in favor of him at 22:30pm—the time at which the quick counts committee decided to make the results public. Notably, models 2 and 3 have a smaller variance than models 1 and 4. Furthermore, model 1 underestimates the proportion of votes, possibly because strata with few polling stations have a bigger impact than on the other models. However, as seen in the right panel of Figure 1, all four models close in on the true value at approximately the same rate.

The real advantage of the models in which the proportion of votes is forced to be an element of Δ^{J-1} is in the estimation of the voter turnout, which can be seen in Figure 2. Inspecting the right panel, we see that the new model (both with and without a complex covariance structure) produces smaller estimates, whereas the original model in both its versions overestimates the turnout. Notably, the modification by Di-Luvi et al. (2018) does a good job at reducing the overestimation, but the new model completely addresses this issue. This is even more clear in the right panel, where we see that, during the early hours of election day—when multiple strata are missing and so samples from the prior are generated—the new model produces considerably shorter credible intervals.

Finally, Figure 3 shows the time required to generate samples from each model. Notably, the model with an arbitrary covariance structure requires a lot of time to mix, even for small sample sizes. Interestingly, the new model with a simple covariance structure requires about the same time as the model of Di-Luvi et al. (2018) with an MCMC sampler. However, the former has the advantage of treating the proportions as elements of the simplex, thereby

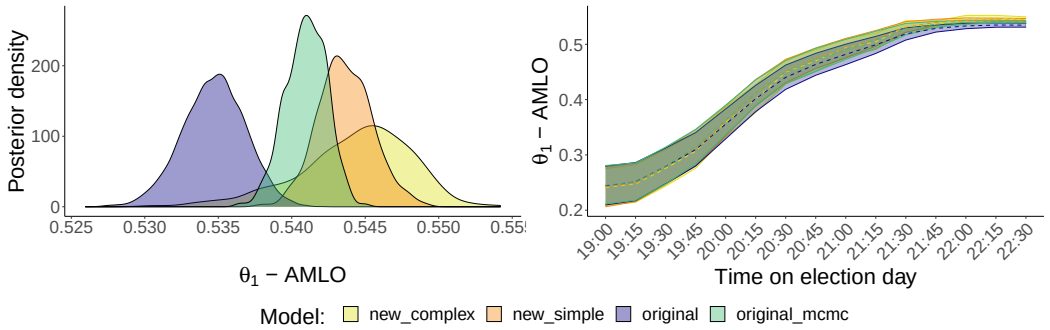

Figure 1: The left panel shows the posterior distribution of the proportion of votes in favor of AMLO at 22:30pm, with models by color. The right panel shows the mean and 0.025 and 0.975 quantiles (giving a 95% posterior credible interval) for different times. `new_complex` refers to model 4, `new_simple` to model 3, `original` to model 1, and `original_mcmc` to model 2.

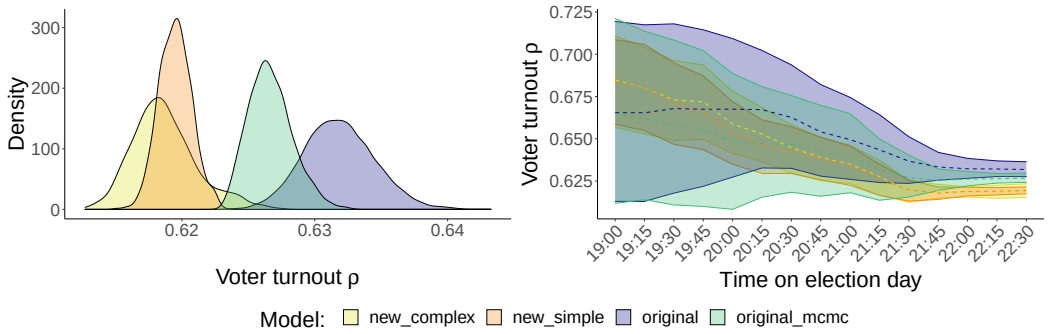

Figure 2: The left panel shows the posterior distribution of the voter turnout at 22:30pm, with models by color. The right panel shows the mean and 0.025 and 0.975 quantiles (giving a 95% posterior credible interval) for different times. `new_complex` refers to model 4, `new_simple` to model 3, `original` to model 1, and `original_mcmc` to model 2.

resulting in better estimations of the voter turnout. It should be noted that using MCMC sampling has a clear impact in computation cost: the model with tractable posterior requires $\mathcal{O}(1)$ time to train, whereas models 2 and 3 become more expensive as sample size increases.

5 Conclusion

In this work, we proposed a new general-purpose Bayesian model for estimating the results of Mexican quick counts that *(i)* addresses the known issue of previous Bayesian models overestimating the voter turnout by considering the proportion of votes as an element of the probabilistic simplex; and *(ii)* can be designed to account for arbitrary covariance structures. We also discussed uninformative prior distributions for both the vector of proportions of votes and covariance matrix and developed Stan code to sample from the posterior of the model.

The new model successfully reduces the voter turnout overestimation even if a trivial covariance structure is assumed. In this regard, we also observed that working with a non-trivial covariance structure results in very slow mixing times and no noticeable gains in accuracy (on top of the benefits that result from working on the simplex). A possible avenue for future work would be considering special decompositions of the covariance structure that give more flexibility than the trivial structure but are not as expensive as the arbitrary

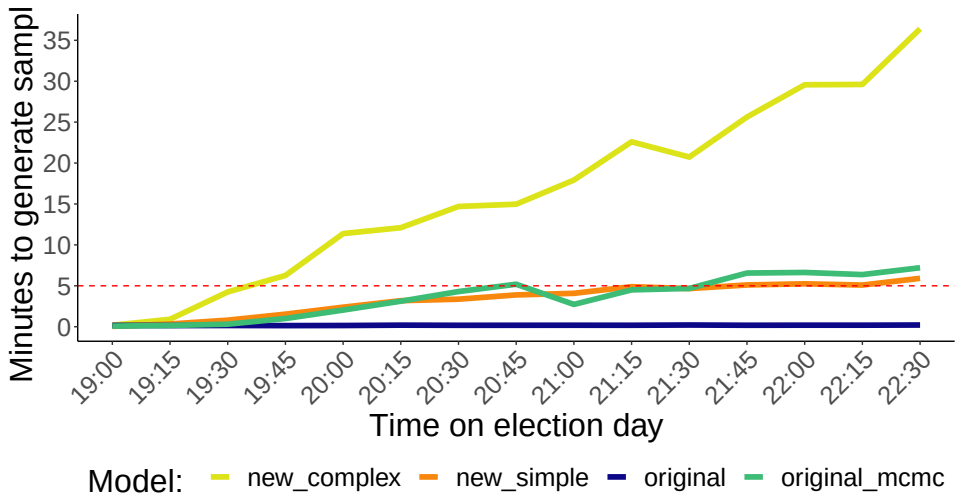

Figure 3: Time in minutes necessary to generate the posterior samples for each model as a function of time. The red dotted line indicates the 5 minute limit for model training. `new_complex` refers to model 4, `new_simple` to model 3, `original` to model 1, and `original_mcmc` to model 2.

structure.

Some of the densities in Figures 1 and 2 have their masses concentrated in very different parts of $[0, 1]$. It should be noted that it is not easy to compare these densities with the known official results of the election: during election day, polling stations officers report the votes after their first count so as to get information to the quick counts committee as early as possible. However, the officers then do a recount, and some votes are subsequently audited in the week after the election. A fair way to compare the models would be to obtain the results from all polling stations, generate multiple sample according to the sampling design used on election day, and train each model in each sample. It would then be possible to compare the errors of each model. However, to the best of our knowledge such a data base is not publicly available. As for the densities in Figures 1 and 2, we believe the discrepancy might be due to the small number of posterior samples generated.

Another line of research is to design faster schemes for model learning. As seen in Figure 3, even the models with a simple covariance structure require over 5 minutes to generate posterior samples when the sample size is relatively large. One possibility would be to use sequential Monte Carlo (Stewart and McCarty Jr., 1992; Gordon et al., 1993; Kitagawa, 1996) or variational inference (Jordan et al., 1999; Wainwright and Jordan, 2008).

References

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982.

M. Anzarut, L. F. González, and M. T. Ortiz. A heavy-tailed multilevel mixture model for the quick count in the mexican elections of 2018. In *National Statistics Forum (FNE) and Latin-American Congress of Statistical Societies (CLATSE)*, pages 1–13. Springer, 2018.

L. E. N. Barajas. Los modelos de estimación: algunas propuestas robustas, 2018.

J. A. Carrera Barroso. Quick counts and exit polls in the mexican elections. *Polis*, 15(2):127–166, 2019.

S. F. J. Cerrillo and L. E. N. Barajas. The veracruz 2016 quick count: Statistical and logistic aspects. *Revista Mexicana de Estudios Electorales*, 3(21), 2019.

G. C. Di-Luvi, G. Orantes-Jordan, and M. Mendoza. Statistics in the 2018 mexican general election quick counts. *Laberintos & Infinitoss*, 48:29–37, 2018.

N. J. Gordon, D. J. Salmond, and A. F. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 107–113. IET, 1993.

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

G. Kitagawa. Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of computational and graphical statistics*, 5(1):1–25, 1996.

D. Lewandowski, D. Kurowicka, and H. Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.

A. Y. Lo. On a class of bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, pages 351–357, 1984.

M. Mendoza and L. E. Nieto-Barajas. Quick counts in the mexican presidential elections: A bayesian approach. *Electoral Studies*, 43:124–132, 2016.

Stan Development Team. RStan: the R interface to Stan, 2020. URL `http://mc-stan.org/`. R package version 2.21.2.

L. Stewart and P. McCarty Jr. Use of Bayesian belief networks to fuse continuous and discrete information for target recognition, tracking, and situation assessment. In V. Libby and I. Kadar, editors, *Signal Processing, Sensor Fusion, and Target Recognition*, volume 1699, pages 177 – 185. International Society for Optics and Photonics, SPIE, 1992. doi: 10.1117/12.138224. URL `https://doi.org/10.1117/12.138224`.

M. J. Wainwright and M. I. Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.