Submitted to the Annals of Applied Statistics

SEMI-PARAMETRIC COVARIATE-MODULATED LOCAL FALSE DISCOVERY RATE FOR GENOME-WIDE ASSOCIATION STUDIES

By Rong W. Zablocki[†,‡],Richard A. Levine[†] Andrew J. Schork[§] Shujing Xu[§] Yunpeng Wang[¶] Chun C. Fan[§] and Wesley K. Thompson[§,‖,**,*]

San Diego State University[†], Claremont Graduate University[‡], University of California at San Diego[§], University of Oslo, Norway[¶], Institute of Biological Psychiatry[‖], and The Lundbeck Foundation Initiative for Integrative Psychiatric Research[**]

1. Supplement A: Conditional Posteriors and Gibbs Sampling Algorithm.

1.1. *Conditional posterior densities.* The full conditional distributions for the unknown parameters may be obtained as follows. Throughout, $f(\cdot|...)$ denotes the kernel probability density of a parameter conditioned on all other parameters and the data. First,

$$f(\boldsymbol{\alpha_{m\cdot}}|...) \propto \exp\left\{-\frac{1}{2\tau_m^2}\left(\boldsymbol{\alpha_{m(2:K)}}\boldsymbol{\Omega^*}\boldsymbol{\alpha_{m(2:K)}}^T + \sum_{j=0;j\neq m}^{M}\left[\boldsymbol{\alpha_{j(2:K)}}\boldsymbol{\Omega^*}\boldsymbol{\alpha_{j(2:K)}}^T\right]\right)\right\} \prod_{i;\delta_i=1}\prod_{k=1}^{K}\left[\left\{\frac{\exp(x_{im}\alpha_{mk}+\boldsymbol{x_{i(-m)}^T}\boldsymbol{\alpha_{(-m)k}})}{\sum_{l=1}^{K}\exp(x_{im}\alpha_{ml}+\boldsymbol{x_{i(-m)}^T}\boldsymbol{\alpha_{(-m)l}})}\right\}^{I(\eta_i=k)}\right]. \tag{1.1}$$

The last part of (1.1) is equivalent to $\prod_{i;\delta_i=1}\prod_{k=1}^{K}\left[\left\{\frac{\exp(\boldsymbol{x_i^T}\boldsymbol{\alpha_{\cdot k}})}{\sum_{l=1}^{K}\exp(\boldsymbol{x_i^T}\boldsymbol{\alpha_{\cdot l}})}\right\}^{I(\eta_i=k)}\right]$, except that (1.1) separates the terms of the m^{th} covariate from all the other covariates (denoted by a $-m$ subscript).

[*] To whom correspondence should be addressed.

Second,

$$\begin{aligned}
&f(\tau_m^2|...) \propto (\tau_m^2)^{-\frac{K-1}{2}} \exp\left\{-\frac{1}{2\tau_m^2}\boldsymbol{\alpha_{m(2:K)}}\boldsymbol{\Omega^*}\boldsymbol{\alpha_{m(2:K)}}^T\right\}(\tau_m^2)^{(-\frac{\nu}{2}-1)} \exp(-\frac{\frac{\nu}{a_m}}{\tau_m^2})\\
&\sim \text{Inverse Gamma}\left(\frac{K+\nu-1}{2}, \frac{\boldsymbol{\alpha_{m(2:K)}}\boldsymbol{\Omega^*}\boldsymbol{\alpha_{m(2:K)}}^T}{2} + \frac{\nu}{a_m}\right).
\end{aligned} \tag{1.2}$$

Third,

$$\begin{aligned}
f(a_m|...) &\propto a_m^{-\frac{\nu}{2}} \exp(-\frac{\frac{\nu}{a_m}}{\tau_m^2}) a_m^{-\frac{1}{2}-1} \exp(-\frac{\frac{1}{A^2}}{a_m})\\
&\sim \text{Inverse Gamma}\left(\frac{\nu+1}{2}, \frac{\nu}{\tau_m^2} + \frac{1}{A^2}\right).
\end{aligned} \tag{1.3}$$

Fourth,

$$\begin{aligned}
f(\sigma_0^2|...) &\propto (\sigma_0^2)^{-(\frac{N_0}{2}+a_0)-1} \exp\left\{-\frac{1}{\sigma_0^2}\left(\frac{\boldsymbol{z_0}^T\boldsymbol{z_0}}{2} + b_0\right)\right\}\\
&\sim \text{Inverse Gamma}\left(\frac{N_0}{2} + a_0,\ \frac{\boldsymbol{z_0}^T\boldsymbol{z_0}}{2} + b_0\right).
\end{aligned} \tag{1.4}$$

Fifth,

$$f(\boldsymbol{\gamma}|...) \propto \prod_{i=1}^{N}\left[\left\{\frac{\exp(\boldsymbol{x_i}^T\boldsymbol{\gamma})}{1+\exp(\boldsymbol{x_i}^T\boldsymbol{\gamma})}\right\}^{\delta_i}\left\{\frac{1}{1+\exp(\boldsymbol{x_i}^T\boldsymbol{\gamma})}\right\}^{1-\delta_i}\right]\exp\left(-\frac{\boldsymbol{\gamma}\Sigma_{\gamma}^{-1}\boldsymbol{\gamma}}{2}\right). \tag{1.5}$$

The distributions in (1.1) and (1.5) do not take any standard distributional form.

1.2. *Sampling Scheme.* In GWAS, each SNP is coded as a count of the number of reference alleles (i.e., 0, 1, or 2). Since choice of reference allele is essentially random with respect to the outcome, the resulting distribution of z-scores is modeled as symmetric around zero. It is straightforward to allow for asymmetry if necessary in other applications of cmfdr. In the present case, in order to simplify the implementation, we fold both the null and non-null distributions at zero. Hence, the null distribution becomes a folded normal distribution with location 0 and scale σ_0^2 , and the non-null distribution is constructed on the absolute value of z-scores. The parameters $\boldsymbol{\alpha}$, $\boldsymbol{\tau}^2$, $\boldsymbol{a}$, $\boldsymbol{\gamma}$ and σ_0^2 are sampled in turn from their full conditional distributions via a Gibbs sampler. At each iteration, σ_0^2 can be sampled directly from its posterior inverse gamma distribution; τ_m^2 and a_m can be sampled from their respective posterior inverse gamma distributions recursively from $m = 0$ to M. The matrix parameter $\boldsymbol{\alpha}$ is generated through compoments $\boldsymbol{\alpha_{m\cdot}}$. At each iteration, $\boldsymbol{\alpha_{m\cdot}}$ may be updated recursively based on (1.1) for $m = 0, 1, 2, ..., M$. While updating $\boldsymbol{\alpha_{m\cdot}}$, all the other $\boldsymbol{\alpha_{-m\cdot}}$ are treated as constant. This is the reason that formula (1.1) separates terms with and without m. Since both $\boldsymbol{\alpha_{m\cdot}}$ and $\boldsymbol{\gamma}$ do not take standard form, variates are generated using multiple-try Metropolis-Hastings (MTMH) samplers (Givens and Hoeting, 2005). Multivariate-t candidate distributions are used in the MTMH sampler with parameters obtained by maximizing $\boldsymbol{\alpha_{m\cdot}}$ and $\boldsymbol{\gamma}$ over (1.1) and (1.5) respectively. That is, the maximizers and the corresponding negative inverse Hessian of (1.1) and (1.5) are used as parameters in the MTMH routine.

Two indicators, $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$, also require updating during each Gibbs iteration. At the global level for each test, we update δ_i to be either 1 or 0 based on the probabilities

$$\frac{\exp(\boldsymbol{x_i}^T\boldsymbol{\gamma})f_1(|z_i||\boldsymbol{\alpha}, \boldsymbol{x_i})}{\exp(\boldsymbol{x_i}^T\boldsymbol{\gamma})f_1(|z_i||\boldsymbol{\alpha}, \boldsymbol{x_i}) + f_0(|z_i||\sigma_0^2)}, \text{ and}$$

$$\frac{f_0(|z_i||\sigma_0^2)}{\exp(\boldsymbol{x_i}^T\boldsymbol{\gamma})f_1(|z_i||\boldsymbol{\alpha}, \boldsymbol{x_i}) + f_0(|z_i||\sigma_0^2)}.$$

The indicator $\boldsymbol{\delta}$ may be initialized by taking an upper percentile (e.g., upper 5%) of the $|z_i|$ scores as 1 and 0 otherwise. At the local level, for each non-null test, we update η_i to be one of the (1, 2,..., K) based on the K probabilities

$$\begin{aligned}
&\frac{\exp(\boldsymbol{x}_{i,\delta_i=1}^{T}\boldsymbol{\alpha}_{\boldsymbol{\cdot 1}})g_1(|z_{i,\delta_i=1}|)}{\sum_{k=1}^{K}[\exp(\boldsymbol{x}_{i,\delta_i=1}^{T}\boldsymbol{\alpha}_{\boldsymbol{\cdot k}})g_k(|z_{i,\delta_i=1}|)]},\\
&\frac{\exp(\boldsymbol{x}_{i,\delta_i=1}^{T}\boldsymbol{\alpha}_{\boldsymbol{\cdot 2}})g_2(|z_{i,\delta_i=1}|)}{\sum_{k=1}^{K}[\exp(\boldsymbol{x}_{i,\delta_i=1}^{T}\boldsymbol{\alpha}_{\boldsymbol{\cdot k}})g_k(|z_{i,\delta_i=1}|)]},\\
&\vdots\\
&\frac{\exp(\boldsymbol{x}_{i,\delta_i=1}^{T}\boldsymbol{\alpha}_{\boldsymbol{\cdot K}})g_K(|z_{i,\delta_i=1}|)}{\sum_{k=1}^{K}[\exp(\boldsymbol{x}_{i,\delta_i=1}^{T}\boldsymbol{\alpha}_{\boldsymbol{\cdot k}})g_k(|z_{i,\delta_i=1}|)]}.
\end{aligned}$$

If we have L draws $\{(\boldsymbol{\alpha}^{(l)}, \boldsymbol{\tau}^{2(l)}, \boldsymbol{a}^{(l)}, \boldsymbol{\gamma}^{(l)}, \sigma_0^{2(l)}) : \ 1 \leq l \leq L\}$ from the Gibbs sampler, then for each draw l,

$$\text{cmfdr}^{(l)}(z_i) = \frac{\pi_0(\boldsymbol{x}_i|\boldsymbol{\gamma}^{(l)})f_0(|z_i||\sigma_0^{2(l)})}{\pi_0(\boldsymbol{x}_i|\boldsymbol{\gamma}^{(l)})f_0(|z_i||\sigma_0^{2(l)}) + \pi_1(\boldsymbol{x}_i|\boldsymbol{\gamma}^{(l)})f_1(|z_i||\boldsymbol{\alpha}^{(l)}, \boldsymbol{x}_i)}.$$

Hence, an *a posteriori* estimate of cmfdr(z_i) for each test can be obtained. Due to the symmetry of f_1 and f_0, if two tests have the same covariates and same absolute value (opposite sign) of z-score, their cmfdr values will be the same. On the other hand, if two tests have identical z-scores but different covariates, their cmfdr values will be different. The algorithm has been implemented in the R statistical package (R Core Team, 2016). The simulations were run on a cluster with 27 Dell PowerEdge C1100 nodes. Each node contains 75 gigabytes (GB) of RAM and dual Intel Xeon X5650 2.66GHz processors with 6 cores for a total of 2025 GB of RAM and 324 cores. A MCMC chain of 18,000 iterations with 50,000 cases took about 20 hours for the semi-parametric cmfdr and about 12 hours for fdr. As for the real data application, all models were run on a Linux Intel Xeon E5-2660 2.20GHz processor with 20 cores for a total of 400 GB of memory. An example of computational resource allocation for a MCMC chain of 23,000 iterations with 74,800 SNPs is shown in Table 1. There is no surprise that the semi-parametric model with covariates is more computationally expensive.

Table 1
Computational resource allocation on three different models.

	semi-parametric cmfdr	gamma cmfdr	fdr
Virtual memory (megabyte)	507	459	495
Physical memory (megabyte)	261	218	249
Time to finish (hours)	55	16	16

2. Supplement B: KEGG *homo sapiens* pathways with ALIGATOR *p*-values from three models (full list).

Table 2: KEGG *homo sapiens* pathways with ALIGATOR *p*-values from three models (full list).

Pathway	*p*-values (semi-parametric)	*p*-values (gamma)	*p*-values (fdr)
Axon guidance	6.00E-04	0.002	0.2046
Herpes simplex infection	8.00E-04	0.0268	1
Osteoclast differentiation	0.0062	0.0192	1
Pentose phosphate pathway	0.0096	0.5206	1
Tuberculosis	0.01	0.0068	0.132
Leishmaniasis	0.0162	0.0946	1
Antigen processing and presentation	0.022	0.096	1
Taste transduction	0.033	1	1
Cytokine-cytokine receptor interaction	0.037	0.0378	1
Cell adhesion molecules (CAMs)	0.0446	0.131	1
Calcium signaling pathway	0.0506	0.1038	0.964
HTLV-I infection	0.0554	0.2448	1
Oocyte meiosis	0.0622	0.8218	0.53
Inflammatory mediator regulation of TRP channels	0.0632	0.1076	1
Prion diseases	0.0642	0.0066	1
TNF signaling pathway	0.0654	0.1124	1
NF-kappa B signaling pathway	0.0672	0.0932	1
Alzheimer's disease	0.079	1	1
Amoebiasis	0.0832	0.2278	1
Pyruvate metabolism	0.0874	0.086	1
Amyotrophic lateral sclerosis (ALS)	0.0892	1	1
Mucin type O-Glycan biosynthesis	0.0902	0.0366	1
Natural killer cell mediated cytotoxicity	0.0906	0.625	1
Circadian rhythm	0.0946	0.096	1
Colorectal cancer	0.0966	0.0354	1
NOD-like receptor signaling pathway	0.0996	0.0946	1
Cocaine addiction	0.102	0.608	0.7526
Apoptosis	0.1094	0.4852	1
T cell receptor signaling pathway	0.1172	0.5708	1
Measles	0.1208	0.512	1

Continued on next page

Table 2 – *Continued from previous page*

Pathway	*p*-values (semi-parametric)	*p*-values (gamma)	*p*-values (fdr)
Hippo signaling pathway	0.121	0.2616	0.7768
MAPK signaling pathway	0.1214	0.1802	0.4044
Salmonella infection	0.1288	0.0486	1
Glycine, serine and threonine metabolism	0.1292	0.0288	0.512
B cell receptor signaling pathway	0.13	0.3578	1
Adipocytokine signaling pathway	0.1312	0.309	1
TGF-beta signaling pathway	0.1342	0.0946	1
Circadian entrainment	0.1394	0.0632	0.7466
Neurotrophin signaling pathway	0.1498	0.8606	0.5142
Influenza A	0.1518	0.3702	1
Steroid hormone biosynthesis	0.1558	1	1
Melanogenesis	0.1582	0.0292	1
Amphetamine addiction	0.1644	0.0952	0.7156
Malaria	0.173	0.3094	1
African trypanosomiasis	0.173	0.3094	1
Cysteine and methionine metabolism	0.174	1	1
Inflammatory bowel disease (IBD)	0.1746	1	1
Fructose and mannose metabolism	0.1748	1	1
Dorso-ventral axis formation	0.1766	0.0946	1
RIG-I-like receptor signaling pathway	0.1774	1	1
Cytosolic DNA-sensing pathway	0.1774	1	1
Sphingolipid metabolism	0.182	1	1
Parkinson's disease	0.1824	1	0.8924
Cell cycle	0.19	0.4334	0.5142
Butanoate metabolism	0.2126	0.5142	1
Vasopressin-regulated water reabsorption	0.2214	0.096	1
Bacterial invasion of epithelial cells	0.2224	0.5174	1
Hepatitis B	0.23	0.0686	1
cGMP-PKG signaling pathway	0.2406	0.0804	0.716
ErbB signaling pathway	0.2652	0.4032	1
Toxoplasmosis	0.2832	0.3702	1
Shigellosis	0.2972	0.0486	1
Thyroid hormone synthesis	0.299	0.232	0.784
Wnt signaling pathway	0.3252	0.8844	1
Notch signaling pathway	0.331	1	1
Basal cell carcinoma	0.3372	0.5162	1
Pathways in cancer	0.3376	0.3212	1
Glycosphingolipid biosynthesis - globo series	0.3394	1	1
Regulation of actin cytoskeleton	0.382	0.3706	1
VEGF signaling pathway	0.4154	0.718	1
Leukocyte transendothelial migration	0.4224	0.2996	1
Purine metabolism	0.4236	1	0.9564
Bladder cancer	0.4426	0.4118	1
Arachidonic acid metabolism	0.4488	0.677	0.7718
Insulin signaling pathway	0.4504	0.313	1

Continued on next page

Table 2 – *Continued from previous page*

Pathway	*p*-values (semi-parametric)	*p*-values (gamma)	*p*-values (fdr)
Glutamatergic synapse	0.4554	0.9532	0.7688
cAMP signaling pathway	0.4558	0.0974	0.9612
AMPK signaling pathway	0.4568	0.2242	1
Estrogen signaling pathway	0.459	0.1794	1
Long-term potentiation	0.4678	0.4244	0.8624
Fc gamma R-mediated phagocytosis	0.4682	0.6288	0.6598
Oxytocin signaling pathway	0.4716	0.3274	0.8514
Non-alcoholic fatty liver disease (NAFLD)	0.479	1	1
Thyroid cancer	0.4884	0.7138	1
Chemokine signaling pathway	0.5262	0.7434	1
Adherens junction	0.5422	0.3858	1
Cholinergic synapse	0.5538	0.4278	0.5386
Retrograde endocannabinoid signaling	0.5574	0.5982	1
Gap junction	0.5602	0.7518	0.9534
Vascular smooth muscle contraction	0.5624	0.6296	1
HIF-1 signaling pathway	0.5664	0.3702	1
PI3K-Akt signaling pathway	0.5668	0.056	0.7768
Toll-like receptor signaling pathway	0.5708	0.1398	1
Inositol phosphate metabolism	0.5916	1	1
Phosphatidylinositol signaling system	0.5916	1	1
Chronic myeloid leukemia	0.6112	0.805	1
Chagas disease (American trypanosomiasis)	0.6184	0.0666	1
Lysine degradation	0.6206	1	1
Focal adhesion	0.6422	0.5276	1
Glycosphingolipid biosynthesis - lacto and neolacto series	0.6526	1	1
ECM-receptor interaction	0.6582	1	1
Renal cell carcinoma	0.66	0.801	1
Prostate cancer	0.6608	0.2842	1
Proteoglycans in cancer	0.6744	0.5572	1
Hedgehog signaling pathway	0.6994	0.7662	1
Morphine addiction	0.7038	1	1
Insulin secretion	0.708	0.801	1
Rap1 signaling pathway	0.7166	0.763	0.9928
Thyroid hormone signaling pathway	0.7388	0.6332	1
Pertussis	0.7656	0.6326	1
Platelet activation	0.7762	0.4428	1
Jak-STAT signaling pathway	0.7918	1	1
Signaling pathways regulating pluripotency of stem cells	0.7926	0.6508	1
Epstein-Barr virus infection	0.7936	1	1
Type II diabetes mellitus	0.794	0.6974	1
GnRH signaling pathway	0.8004	0.2552	1
mTOR signaling pathway	0.8034	0.5968	1
Adrenergic signaling in cardiomyocytes	0.8068	0.0134	0.9956
GABAergic synapse	0.8304	0.7428	1
beta-Alanine metabolism	0.8356	0.6766	0.5264

Continued on next page

Table 2 – *Continued from previous page*

Pathway	p-values (semi-parametric)	p-values (gamma)	p-values (fdr)
Drug metabolism - other enzymes	0.8356	0.6766	0.5264
Fc epsilon RI signaling pathway	0.8394	0.5604	1
Acute myeloid leukemia	0.8416	0.5968	1
Non-small cell lung cancer	0.8416	0.5968	1
Pancreatic cancer	0.8416	0.5968	1
Endometrial cancer	0.8416	0.5968	1
Prolactin signaling pathway	0.8442	0.5606	1
MicroRNAs in cancer	0.8496	0.1518	0.3328
Pantothenate and CoA biosynthesis	0.864	0.7748	0.5264
Tight junction	0.8834	0.17	1
Long-term depression	0.8862	0.4218	0.5324
Dopaminergic synapse	0.8868	0.015	0.925
Glioma	0.9018	0.5968	1
Progesterone-mediated oocyte maturation	0.9018	0.7476	1
FoxO signaling pathway	0.9022	0.7088	1
Serotonergic synapse	0.9106	0.9062	1
Hepatitis C	0.9122	0.1406	1
Pyrimidine metabolism	0.9238	0.9284	0.5336
Dilated cardiomyopathy	0.9456	1	1
Melanoma	0.957	0.7938	1
Alcoholism	0.9636	0.9998	0.0822
Olfactory transduction	0.9852	0.9458	0.9826
Ras signaling pathway	0.9864	0.8614	1
Pathogenic Escherichia coli infection	1	0.092	1
N-Glycan biosynthesis	1	0.1164	0.2634
Pentose and glucuronate interconversions	1	0.3054	1
Epithelial cell signaling in Helicobacter pylori infection	1	0.4482	1
Alanine, aspartate and glutamate metabolism	1	0.5142	1
Aminoacyl-tRNA biosynthesis	1	0.7708	0.8938
alpha-Linolenic acid metabolism	1	1	1
Graft-versus-host disease	1	1	1
Histidine metabolism	1	1	1
Huntington's disease	1	1	1
Hypertrophic cardiomyopathy (HCM)	1	1	1
Intestinal immune network for IgA production	1	1	1
Legionellosis	1	1	1
Linoleic acid metabolism	1	1	1
Lipoic acid metabolism	1	1	1
Lysine biosynthesis	1	1	1
Amino sugar and nucleotide sugar metabolism	1	1	1
Maturity onset diabetes of the young	1	1	1
Metabolism of xenobiotics by cytochrome P450	1	1	1
Mineral absorption	1	1	1
Neuroactive ligand-receptor interaction	1	1	1
Nicotinate and nicotinamide metabolism	1	1	0.7818

Continued on next page

Table 2 – *Continued from previous page*

Pathway	*p*-values (semi-parametric)	*p*-values (gamma)	*p*-values (fdr)
Nitrogen metabolism	1	1	1
One carbon pool by folate	1	1	1
Ovarian steroidogenesis	1	1	1
Oxidative phosphorylation	1	1	1
p53 signaling pathway	1	1	1
Pancreatic secretion	1	1	1
Phenylalanine metabolism	1	1	1
Phenylalanine, tyrosine and tryptophan biosynthesis	1	1	1
Phototransduction	1	1	1
Porphyrin and chlorophyll metabolism	1	1	1
Primary bile acid biosynthesis	1	1	1
Propanoate metabolism	1	1	1
Proximal tubule bicarbonate reclamation	1	1	1
Retinol metabolism	1	1	1
Rheumatoid arthritis	1	1	1
Riboflavin metabolism	1	1	1
Salivary secretion	1	1	1
Selenocompound metabolism	1	1	1
Small cell lung cancer	1	1	1
Staphylococcus aureus infection	1	1	1
Starch and sucrose metabolism	1	1	1
Steroid biosynthesis	1	1	1
Sulfur metabolism	1	1	1
Synaptic vesicle cycle	1	1	1
Arginine and proline metabolism	1	1	1
Synthesis and degradation of ketone bodies	1	1	1
Systemic lupus erythematosus	1	1	1
Taurine and hypotaurine metabolism	1	1	1
Terpenoid backbone biosynthesis	1	1	1
Thiamine metabolism	1	1	1
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	1	1	1
Transcriptional misregulation in cancer	1	1	1
Tryptophan metabolism	1	1	1
Type I diabetes mellitus	1	1	1
Asthma	1	1	1
Tyrosine metabolism	1	1	1
Ubiquinone and other terpenoid-quinone biosynthesis	1	1	1
Valine, leucine and isoleucine biosynthesis	1	1	1
Valine, leucine and isoleucine degradation	1	1	1
Vibrio cholerae infection	1	1	0.7708
Viral carcinogenesis	1	1	1
Viral myocarditis	1	1	1
Autoimmune thyroid disease	1	1	1
Vitamin B6 metabolism	1	1	1

Continued on next page

Table 2 – *Continued from previous page*

Pathway	p-values (semi-parametric)	p-values (gamma)	p-values (fdr)
Vitamin digestion and absorption	1	1	1
Bile secretion	1	1	1
Biotin metabolism	1	1	1
Butirosin and neomycin biosynthesis	1	1	1
Caffeine metabolism	1	1	1
Carbohydrate digestion and absorption	1	1	1
Cardiac muscle contraction	1	1	1
Chemical carcinogenesis	1	1	1
Citrate cycle (TCA cycle)	1	1	1
Complement and coagulation cascades	1	1	1
Cyanoamino acid metabolism	1	1	1
D-Glutamine and D-glutamate metabolism	1	1	1
Drug metabolism - cytochrome P450	1	1	1
Endocrine and other factor-regulated calcium reabsorption	1	1	1
Ether lipid metabolism	1	1	1
Fat digestion and absorption	1	1	1
Fatty acid biosynthesis	1	1	1
Fatty acid degradation	1	1	1
Fatty acid elongation	1	1	1
Folate biosynthesis	1	1	1
Aldosterone-regulated sodium reabsorption	1	1	1
Galactose metabolism	1	1	1
Gastric acid secretion	1	1	0.8974
Glutathione metabolism	1	1	1
Glycerolipid metabolism	1	1	1
Allograft rejection	1	1	1
Glycerophospholipid metabolism	1	1	1
Glycolysis / Gluconeogenesis	1	1	1
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	1	1	1
Glycosaminoglycan degradation	1	1	1
Glycosphingolipid biosynthesis - ganglio series	1	1	1
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	1	1	1
Glyoxylate and dicarboxylate metabolism	1	1	1

3. Supplement C: Identifiability of the mixture model.

First consider the model without covariates. At the global level, the two-component mixture model under the symmetric assumption has marginal density $f(|z|) = (1-\pi_1)f_0(|z|) + \pi_1 f_1(|z|)$, where f_0 is a folded $N(0, \sigma_0^2)$, f_1 is a nonparametric non-null distribution with known location $\mu \geq 0.68$, and

π_1 takes the form of the logistic function $\frac{exp(\gamma)}{1+exp(\gamma)}$. This model is identifiable upon making the center null assumption for f_0 and "zero assumption" of Efron (2007). At the local level, the non-null density is approximated by cubic B-spline density $g_k(|z|)$ and mixing weight c_k such that $f_1(|z|) = \sum_{k=1}^{K} c_k g_k(|z|)$. The mixing weight c_k is a multinomial function $\frac{exp(\boldsymbol{\alpha}_k)}{\sum_{l=1}^{K} exp(\boldsymbol{\alpha}_l)}$ where $\sum_{k=1}^{K} c_k = 1$. According to Inkila (1988) and De Boor et al. (1978), a sequence of cubic B-splines generated over a strictly increasing sequence of knots is uniquely defined. The overall model is thus identifiable in the case of no covariates.

From here, identifiability of our covariate-modulated model follows in an analogous manner to Theorem 1 of Huang and Yao (2012). In particular, the covariates $\boldsymbol{x}$ enter our model through $\boldsymbol{x}^T\boldsymbol{\gamma}$ and $\boldsymbol{x}^T\boldsymbol{\alpha}$ components in the mixing weights π_1 and c_k respectively. The model conditioning on the covariates $\boldsymbol{x}$ is identifiable. Assuming the domain of the covariates $\boldsymbol{x}$ has no isolated points, identifiability of our proposed mixture model, then follows through in an identical manner to the proof of Theorem 1 of Huang and Yao (2012).

4. Supplement D: Convergence Diagnosis Plots.

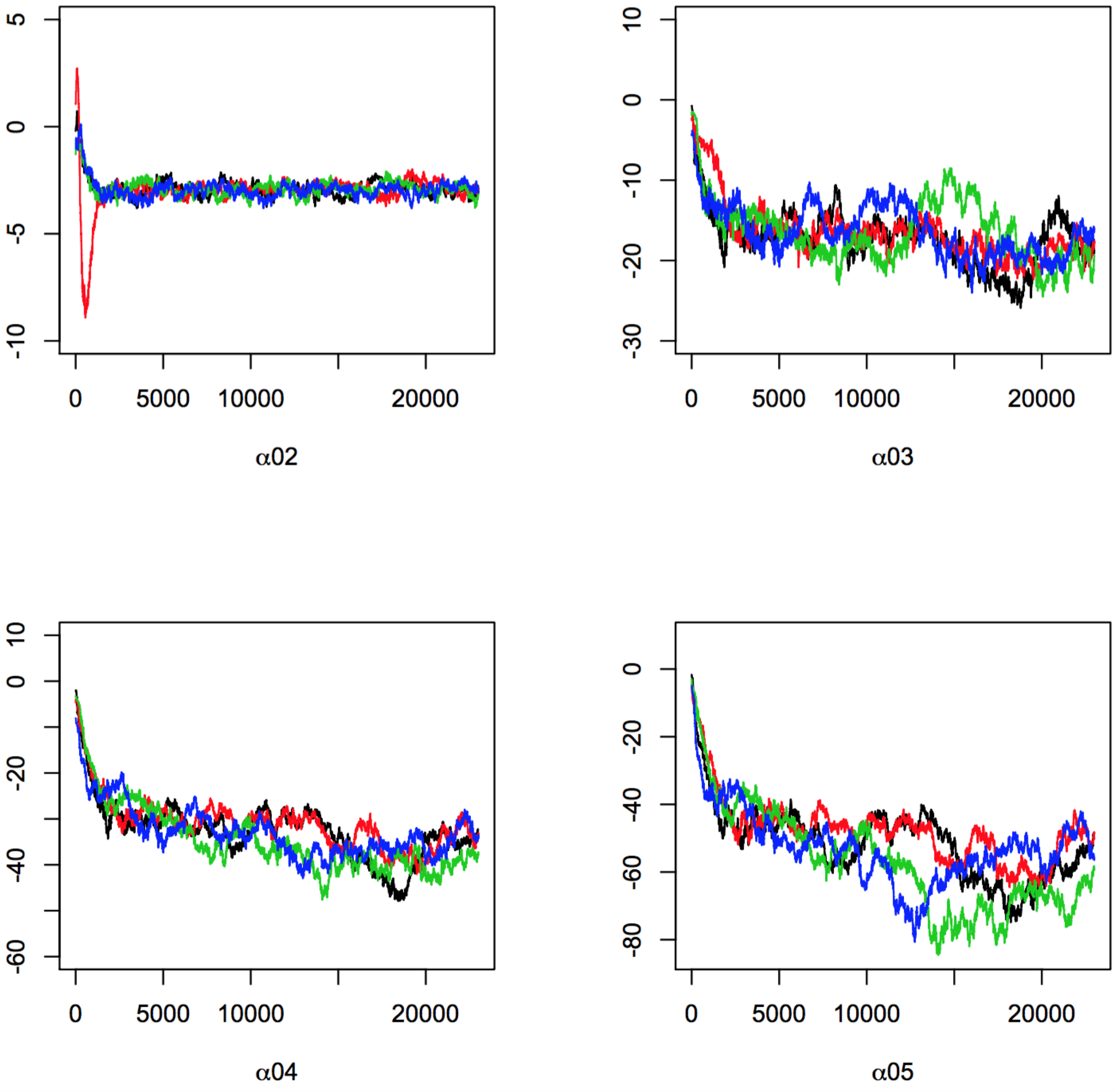

Fig 1: Running mean plot of $\boldsymbol{\alpha}_{\mathbf{0}\cdot}$, 23000 iterations, 4 chains. (note: $\boldsymbol{\alpha}_{01} = 0$).

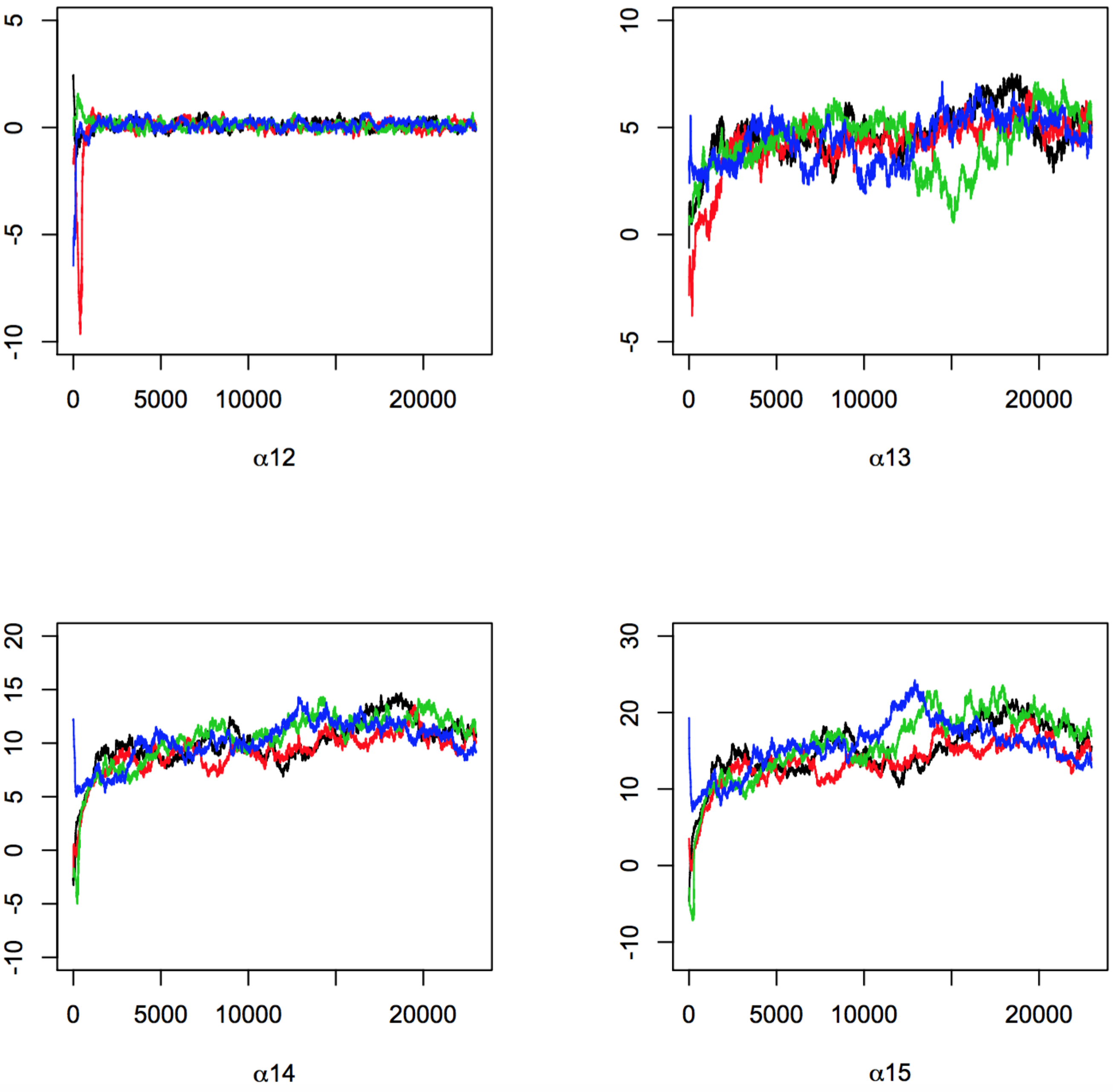

Fig 2: Running mean plot of $\boldsymbol{\alpha_{1\cdot}}$, 23000 iterations, 4 chains. (note: $\boldsymbol{\alpha}_{11} = 0$).

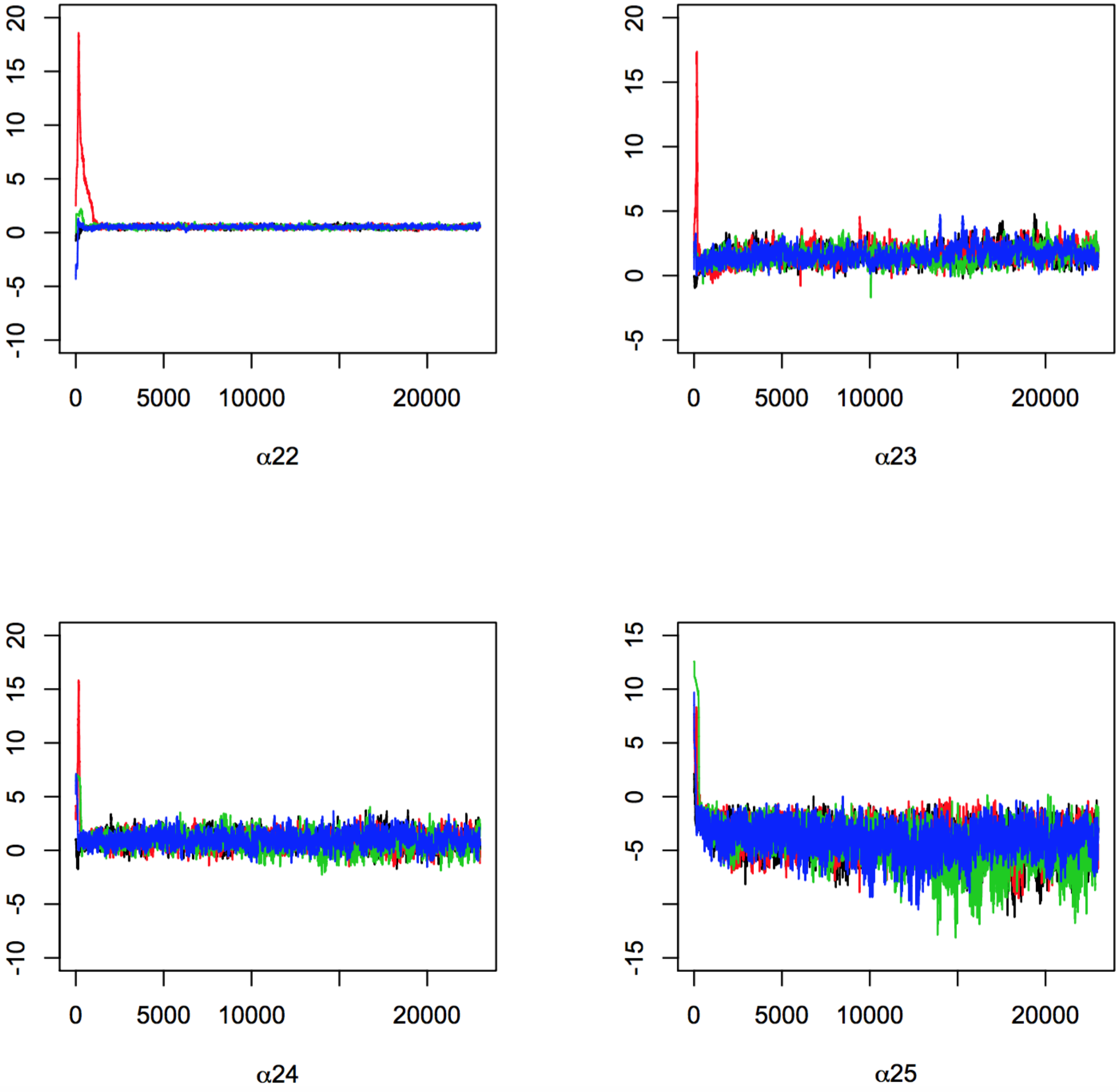

Fig 3: Running mean plot of $\boldsymbol{\alpha_{2\cdot}}$, 23000 iterations, 4 chains. (note: $\boldsymbol{\alpha}_{21} = 0$).

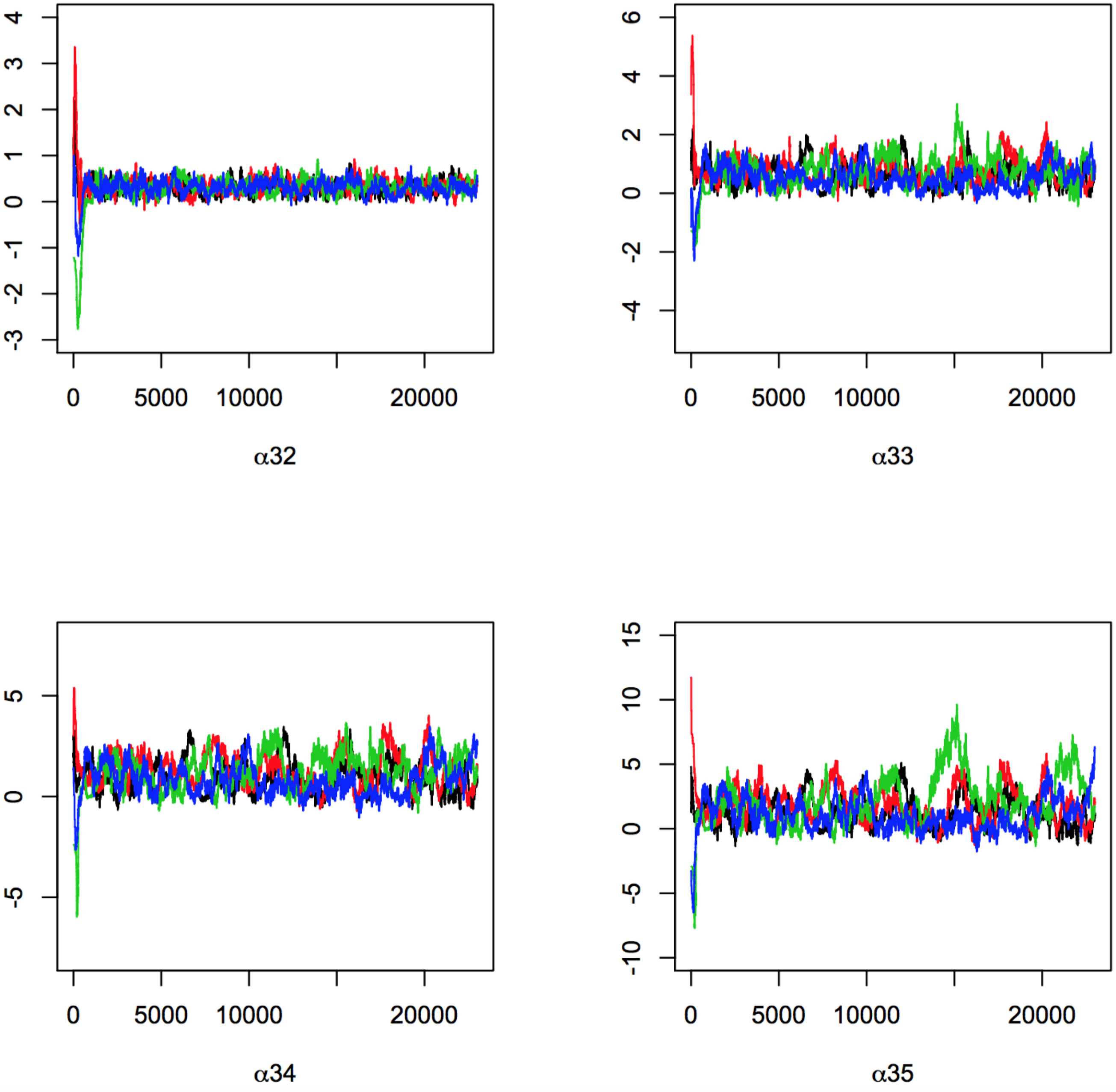

Fig 4: Running mean plot of $\boldsymbol{\alpha}_{\mathbf{3}\cdot}$, 23000 iterations, 4 chains. (note: $\boldsymbol{\alpha}_{31} = 0$).

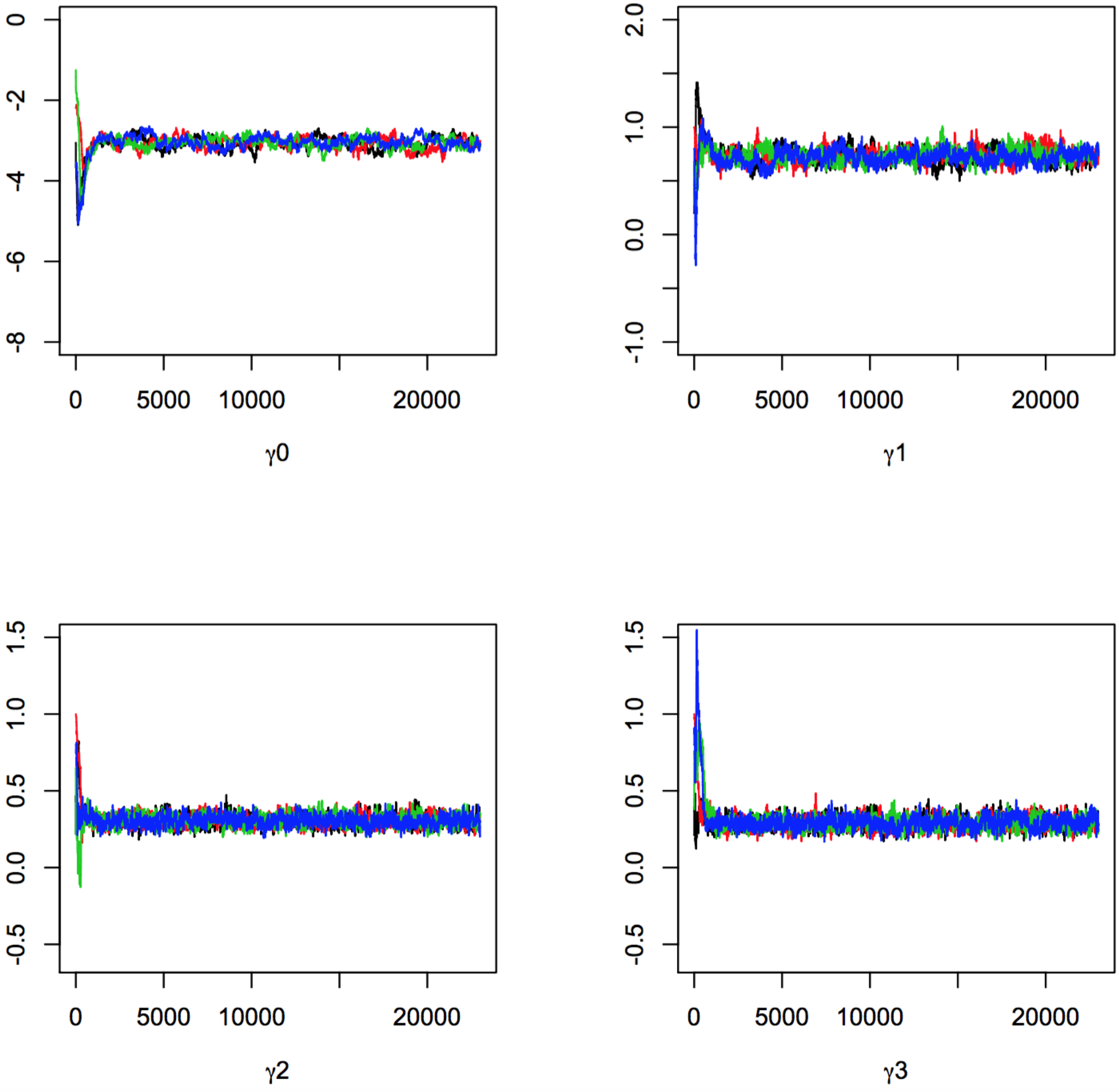

Fig 5: Running mean plot of $\boldsymbol{\gamma}$, 23000 iterations, 4 chains.

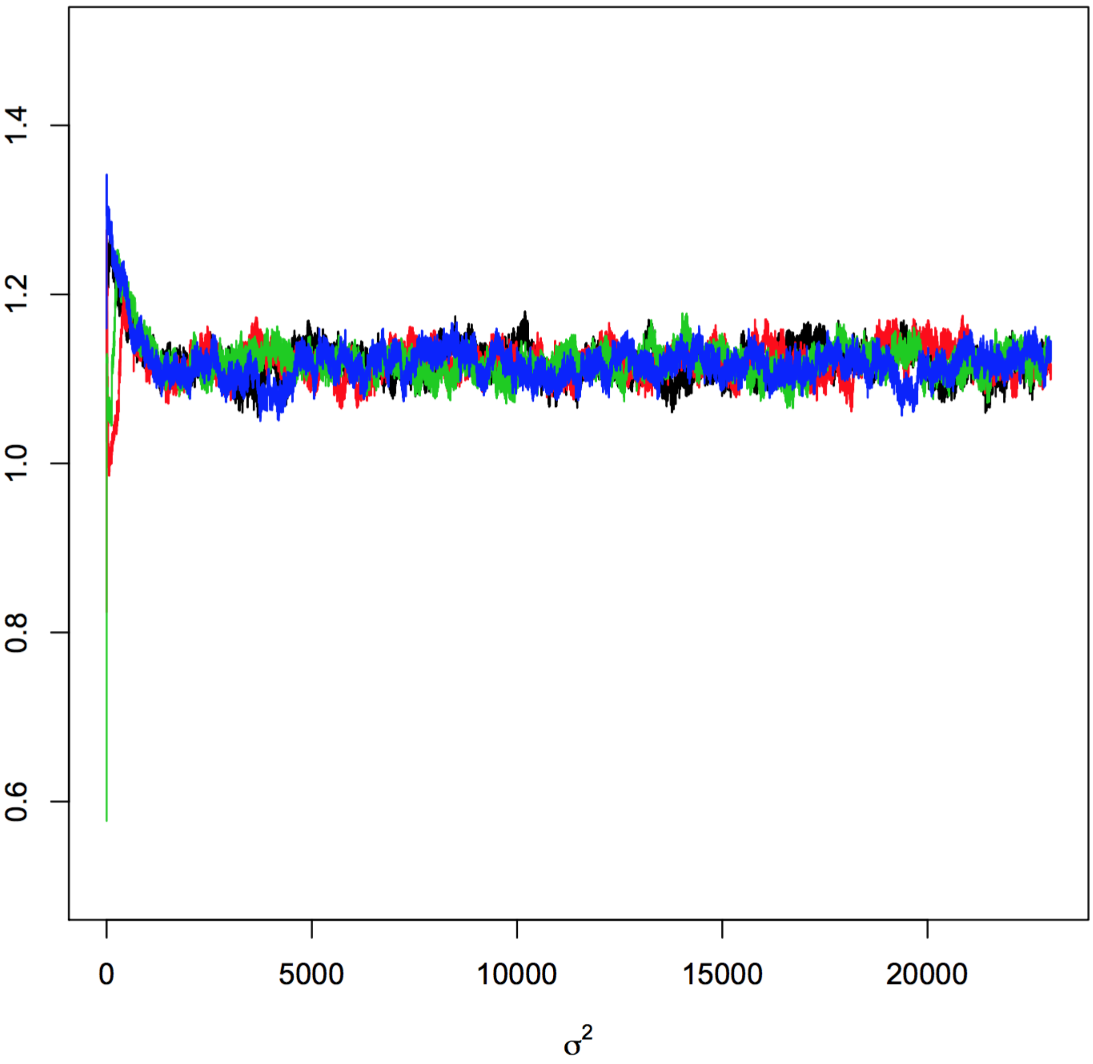

Fig 6: Running mean plot of σ_0^2, 23000 iterations, 4 chains.

References.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C. and De Boor, C. (1978). *A practical guide to splines* **27**. Springer-Verlag New York.

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics* 1351–1377.

Givens, G. H. and Hoeting, J. A. (2005). *Computational statistics* **483**. Wiley Interscience Press.

Huang, M. and Yao, W. (2012). Mixture of regression models with varying mixing proportions: a semiparametric approach. *Journal of the American Statistical Association* **107** 711–724.

Inkila, K. (1988). Bicubic B-spline approximation by least squares. In *XVIth ISPRS Congress, Technical Commission III: Mathematical Analysis of Data* **XXVII Part B3** 281– 287.

R Core Team (2016). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Rong W. Zablocki
Computational Science Research Center
San Diego State University
5500 Campanile Drive
San Diego, CA 92182
USA;
Institute of Mathematical Sciences
Claremont Graduate University
150 E. 10th St.
Claremont, CA 91711
USA

Richard A. Levine
Department of Mathematics and Statistics
San Diego State University
5500 Campanile Drive
San Diego, CA 92182
USA

Andrew J. Schork
Cognitive Sciences Graduate Program
University of California at San Diego
9500 Gilman Drive
La Jolla, CA 92093
USA

Shujing Xu
Department of Psychiatry
University of California at San Diego
9500 Gilman Drive
La Jolla, CA 92093
USA

Yunpeng Wang
Institute of Clinical Medicine
University of Oslo
Oslo, 0424
Norway

Chun C. Fan
Cognitive Sciences Graduate Program
University of California at San Diego
9500 Gilman Drive
La Jolla, CA 92093
USA

Wesley K. Thompson
Institute of Biological Psychiatry
Mental Health Centre Sct. Hans
Mental Health Services Copenhagen
DK-4000
Denmark
Department of Psychiatry
University of California at San Diego
9500 Gilman Drive
La Jolla, CA 92093
USA E-mail: wes.stat@gmail.com